No Harm in Checking:
Using Factual Manipulation Checks to Assess Attentiveness in Experiments

Short title for running header: "Factual Manipulation Checks"

Keywords: manipulation checks; experiments; screeners

John V. Kane
Center for Global Affairs
New York University
Room 437A
15 Barclay Street,
New York, NY 10007
john.kane@nyu.edu

Jason Barabas
Department of Political Science
Stony Brook University
Social & Behavioral Sciences, 7th Floor
Stony Brook, NY 11794
jason.barabas@stonybrook.edu

Acknowledgements: We would like to thank Jennifer Jerit, Yanna Krupnikov, David Nickerson, Scott Clifford, and Christine Peterson for their helpful comments on earlier drafts of this manuscript. Additionally, Michelle lo-Low and Mike Kriner provided outstanding research assistance. Finally, we also appreciate the suggestions that we received from the anonymous reviewers, faculty and graduate students at Stony Brook University, and panelists at the annual conferences of the American Political Science Association and the International Society for Political Psychology.

Abstract

Manipulation checks are often advisable in experimental studies, yet they rarely appear in practice. This lack of usage may stem from fears of distorting treatment effects and uncertainty regarding which type to use (e.g., instructional manipulation checks [IMCs] or assessments of whether stimuli alter a latent independent variable of interest). Here, we first categorize the main variants and argue that *factual manipulation checks* (FMCs)—i.e., objective questions about key elements of the experiment—can identify individual-level attentiveness to experimental information and, as a consequence, better enable researchers to diagnose experimental findings. We then find, through four replication studies, little evidence that FMC placement affects treatment effects, and that placing FMCs immediately post-outcome does not attenuate FMC passage rates. Additionally, FMC and IMC passage rates are only weakly related, suggesting that each technique identifies different sets of attentive subjects. Thus, unlike other methods, FMCs can confirm attentiveness to experimental protocols.

Replication Materials: The data, code, and any additional materials required to replicate all analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: http://dx.doi.org/XXX

WORD COUNT: 9,981

Experiments are a popular method for testing hypotheses within the social sciences. Essential for the validity an experiment is the extent to which a study's participants are actually "treated" (Oppenheimer, Meyvis, and Davidenko 2009). As a precondition for being treated, respondents must (in most instances) first be attentive to treatments. Ensuring that respondents attend to stimuli, however, presents a challenge, particularly given the growth in online surveys and reliance upon subjects from opt-in panels who may be distracted (e.g., Chandler and Shapiro 2016; Clifford and Jerit 2015). Within a study, therefore, being able to accurately gauge receipt of treatment is essential for testing hypotheses and advancing theory.

One technique for assessing treatment receipt is that of a manipulation check (MC). Broadly defined, MCs are "used to check whether the manipulation conducted in an experiment is perceived by the subjects as the experimenter wishes it to be perceived" (Morton and Williams 2010, 108). However, there is much variation in *why* researchers implement MCs. In practice, scholars sometimes report using manipulation checks to determine whether the latent independent variable of interest has been affected by experimental stimuli (Mutz and Pemantle, 2015, 196). More generally, researchers also use MCs to assess respondent *attentiveness* during a study, e.g., via specific questions given to participants to judge whether they are reading carefully (e.g., Anduiza and Galais 2016; Crump, McDonnell, and Gureckis 2013; Maniaci and Rogge 2014). MCs therefore provide researchers with leverage on the nature of their experimental findings. As Mutz (2011, 84–85) contends, "Experiments for which manipulation checks can be considered 'optional' are relatively few and far between," primarily because failing to include an MC "undermines what can be learned from any given study."

Perhaps because of the varying functions of MCs, there is widespread heterogeneity in the *forms* that MCs can take. Moreover, MCs vary in terms of their *placement* within an

experiment, which is important because scholars worry that the placement of MCs might alter treatment effects (Berinsky, Margolis, and Sances 2014; Mutz 2011), thereby undermining the internal validity of an experiment. Thus, while MCs vary because they are tailored for each study, MC usage also varies more generally in terms of both form and placement. This heterogeneity makes it challenging for researchers to ascertain best practices for implementing MCs.

Our study aims to clarify the use of MCs in several ways. Guided by results from a content analysis of experimental research, we first categorize the major variants of MCs as well as their advantages and disadvantages. Ultimately, we present a type of manipulation check that is not well recognized in political science— *factual* MCs (FMCs), which ask respondents objective questions (i.e., questions with correct answers) about a study's content—and suggest that FMCs can serve as useful tools for researchers who employ experiments.

But how should FMCs be implemented? The remaining portion of our study replicates four previous survey experiments, allowing us to investigate whether FMC placement alters treatment effects or creates difficulties for identifying attentive respondents. Overall, the results provide little evidence that placing an FMC immediately *before* an outcome measure distorts treatment effects; any distortions were small in size, inconsistently signed, and rarely attained statistical significance. We also found that placing FMCs *immediately after* the outcome measure avoids distorting treatment effects without incurring any costs in terms of measuring respondents' attentiveness. Finally, having also asked instructional manipulation checks (IMCs) in three of our studies, we note important differences in respondent attentiveness when measured with FMCs versus IMCs. To our knowledge, our study represents the first attempt to explore the

nature, usage, and implementation of manipulation checks. As such, our findings should help improve experimental design and analysis.

MANIPULATION CHECKS IN PRACTICE

In using experiments to test hypotheses, it is crucial that subjects receive treatments. When researchers wish to statistically assess whether treatments were perceived as intended by subjects in the experiment, a common method is a manipulation check (MC) (Morton and Williams 2010). Despite their usefulness, however, many experimental analyses do not utilize MCs (Mutz 2011, 84). As detailed in the Supplemental Appendix, we conducted a content analysis of experimental studies in prominent journals from 2001 to 2015 and found that only 62 articles employed MCs out of 338 (18%). Among those studies that report having used an MC, there is considerable heterogeneity in the type used. In this section, we identify and describe three main categories of manipulation checks.

Subjective Manipulation Checks

A popular form of an MC is what we refer to as a *subjective manipulation check.*[1] Subjective MCs (SMCs) typically ask respondents for their thoughts regarding the independent variable being manipulated by the researcher. For example, in investigating whether a national draft would affect Americans' support for going to war, Horowitz and Levendusky (2011) attempt to manipulate perceptions about military draft reinstatement. Thus, the authors' MC asks respondents to "assess the likelihood that the draft will be reintroduced using a 5-point order response scale ranging from 'very unlikely' to 'very likely"'(2011, 529). Responses to this item are inherently *subjective*; but, if the treatment is efficacious, the treatment group should, on

[1] Indeed, our content analysis found that even though MCs are rare, SMCs are most common, having been used in 36 of these 62 articles (58%).

average, view draft reintroduction as more likely than the control group. Similarly, Dietrich and Winters (2015) are interested in whether foreign aid improves or undermines government legitimacy. The authors thus manipulated the stated source of aid for an HIV/AIDS prevention program, and asked Indian respondents, "How much do you think foreign aid from other countries contributes to the provision of social services in India?" (2015, 167). Response options ranged on a 4-point scale from "nothing" to "a lot." Again, the expectation is that subjects in the treatment condition(s) will have a significantly higher mean on the SMC item than control group subjects. Such results are taken as evidence of treatment receipt. For example, Dietrich and Winters conclude that, "This significant treatment effect is evidence that the experimental manipulations conveyed the intended information about the foreign funding of development interventions and respondents absorbed this information" (2015, 167).

Researchers can therefore glean important information from the results of SMCs. However, a defining characteristic of SMCs is that it is typically not possible for individuals to give a *wrong* answer. For example, if a respondent in the treatment condition of Dietrich and Winters's study had answered in the MC that foreign aid contributes "a little" to the provision of services, the authors cannot infer that this respondent failed to attend to the treatment.[2] Again, responses to SMCs are *subjective* in nature, and this precludes the researcher from identifying individuals who were inattentive to the treatment stimulus. Instead, researchers must compare SMC responses across *groups* rather than across individuals—i.e., researchers can test only whether the treatment group differs, on average, from the control group. Thus, if only an SMC is used, researchers cannot assess individual attentiveness to the experimental stimuli, even though

[2] Similarly, had the individual answered "a lot," the authors cannot infer that the individual paid attention to the treatment because this may have been the respondent's perception without treatment. Most importantly, one cannot argue that there is a single "correct" answer.

inattentiveness is a common problem among respondents who may be unmotivated and engaging in "satisficing" (Krosnick, Narayan, and Smith 1996). Again, attentiveness is a crucial precondition for treatments to be efficacious.[3]

Instructional Manipulation Checks

In contrast to subjective MCs, the *instructional manipulation check* (IMC) is an alternative technique that directly attempts to assess attentiveness (either in experiments or surveys more generally). IMCs typically ask respondents a generic question (e.g., "Which of these [sports] activities do you engage in regularly?"), along with a variety of closed-ended response options (Oppenheimer, Meyvis, and Davidenko 2009; also see Curran 2016; Lovett et al. 2018; Maniaci and Rogge 2014). However, embedded within the prelude to this question are specific instructions to the respondent regarding which answer choices to select and what information to disregard. The logic is that respondents who have little motivation to complete the experiment in earnest will satisfice, giving answers to the generic question while overlooking the embedded instructions. Such individuals, having failed the IMC instructions, are typically regarded as inattentive. For example, Clifford et al. (2015, 1182) used an IMC that asked respondents "which sections of the newspaper they like to read," whereas the preceding instructions directed respondents to select only "classifieds" and "none of the above." Subjects who did not "pass" this IMC were "considered inattentive" and not permitted to complete the survey.

As the example of Clifford et al. (2015) illustrates, IMCs—or, "screeners"—have advantages over subjective MCs: namely, IMCs enable researchers to identify inattentive

[3] While attention is important in most experiments, some interventions may be subconscious (e.g., subliminal priming studies).

individuals. Nevertheless, despite their strengths, IMCs have drawbacks. First, and most importantly, IMCs are unrelated to the experiment being conducted by the researcher. Thus, though they may measure attentiveness to the survey in general, IMCs do not measure attentiveness to stimuli contained within the *experimental portion* of the survey. This is especially problematic insofar as respondents' levels of attentiveness vary throughout the course of a study (Anduiza and Galais 2016) or in surveys with many distractions (Clifford and Jerit 2014); thus, attentiveness to any given IMC may be a poor indicator of attentiveness to the experimental manipulations. Second, individuals who regularly complete surveys online (e.g., MTurk "Workers") may become exceptionally "savvy" at spotting IMCs (see Krupnikov and Levine 2014; Hauser and Schwarz 2016) and/or share such information with other users (Chandler, Mueller, and Paolacci 2014). Thus, unless researchers design unique IMCs for their studies, using previously established IMCs may be ineffective if subjects have become familiar with these IMCs (and/or their general appearance) in advance.

Moreover, because the IMC is essentially a trick question, some researchers have investigated whether using an IMC affects the way respondents answer subsequent questions. In other words, IMCs may not benignly measure attentiveness—rather, they might *affect* attentiveness (Hauser and Schwarz 2015). This poses potential problems, especially if IMCs are employed before the experimental manipulation of interest. If respondents think the researcher is attempting to "trick them," effect estimates may be biased, particularly when examining effects only among IMC passers. Lastly, featuring multiple IMCs (as recommended by Berinsky et al. [2014]) can be both costly and can make presenting experimental results more complicated (e.g., treatment effects for those who passed only one screener, two, three, etc.).

Factual Manipulation Checks

Finally, an MC that combines the strengths of SMCs and IMCs is a *factual manipulation check*. This term is essentially unknown in political science, yet is a conceptually important innovation. Operationally, an FMC is typically a single question that is asked of all participants in the experiment. In a broad sense, FMCs are simply MCs that appear post-treatment and are pertinent to the experimental portion of the survey (like SMCs), but that also have *correct response options* (like IMCs). In short, the question asked of respondents is factual in nature—i.e., a given response is either correct or incorrect (see the supplemental appendix for more on the implicit "correct answer" assumption). For example, Turner (2007) manipulates whether various news stories are attributed to CNN or the Fox News Channel, and thus asks all respondents in the study to indicate "what network produced the stories they viewed." Notice that the question is (1) specifically about the *experimental material*, and (2) any given answer choice is either factually correct or incorrect. Again, these attributes are what differentiate FMCs from other types of MCs, thereby enabling the researcher to identify *respondents who were attentive to the experiment* (see also Broockman and Butler [2017, 5] for a field experiment example). Turner (2007, 449) reports that 92.4% answered this FMC correctly, and that, as is sometimes done by other researchers (Hauser and Schwarz 2015), those who failed the MC were excluded from the analysis.[4]

Factual MCs are also flexible. Namely, they can question respondents about information that is either relevant *or* irrelevant to the treatment stimulus. In some experiments—e.g., Turner (2007) —it may be possible to ask all respondents the same *treatment-relevant* FMC (FMC-TR).

[4] While we acknowledge that some researchers use FMCs to conduct analyses on "passers" only, this practice raises concerns regarding post-treatment bias (e.g., Acharya, Blackwell, and Sen 2016; Montgomery, Nyhan and Torres 2018) and/or unrepresentative samples (Berinsky, Margolis, and Sances 2014, 747–48).

This enables the researcher to not only investigate whether *individuals* responded correctly to the FMC, but also whether experimental *groups* significantly differ in their FMC responses. In other words, responses to FMCs can be analyzed at the individual-level (e.g., to determine whether a given individual was attentive, or to calculate the proportion of attentive individuals in a given condition or the study as a whole), or at the group-level (e.g., by confirming that respondents in the "CNN" condition, for example, answered "CNN" to a significantly greater extent than respondents in the "Fox News Channel" condition).

Yet, sometimes researchers may find it challenging to ask an FMC-TR of individuals in all experimental condition(s). For example, Maoz (2012) presents respondents with a photo of an adult male with digitally manipulated facial features (specifically, the size of the lips and eyes). Here, asking respondents in each experimental condition a factual question specifically about the manipulated material is unlikely to be fruitful (e.g., asking respondents about eye size would not easily permit informative responses).

In such cases, researchers can use an FMC that asks about material in the experimental portion of the survey that was *not* manipulated across conditions—i.e., a treatment-*irrelevant* FMC (FMC-TI). Returning to the example of Maoz (2012), the researcher might ask of the respondents, "What color was the adult male's hair?" In this way, the researcher would be able to identify individuals who were attentive to the experimental portion of the survey (see also Healy and Lenz 2014, 39 fn13). However, since an FMC-TI will have the same correct response across experimental conditions, researchers cannot use an FMC-TI to assess whether the manipulation was perceived. Thus, an FMC-TR has advantages over an FMC-TI because the former measures attentiveness to the content being manipulated, thereby enabling researchers assess respondents'

attentiveness to the experiment and test for significant differences in response patterns across experimental groups.

Summary: Advantages of FMCs over SMCs and IMCs

Factual manipulation checks (FMCs) give researchers a unique ability to evaluate their experimental findings. As illustrated earlier, FMCs enable researchers to better identify which individuals were attentive to experimental material. Further, FMCs (in particular, FMC-TRs) enable researchers to test whether significant differences in response patterns exist between experimental groups (which should be the case given that the FMC-TR asks respondents about information that varies across groups).

FMCs thus have distinct advantages over both SMCs and IMCs. SMCs, like FMCs, can assess significant differences in responses between experimental groups. But, because of their subjective nature, SMC responses cannot determine whether a given individual was attentive to the experimental information. In other words, an SMC does not measure attentiveness to the experiment—it *assumes* it. Alternatively, IMCs, which have gained prominence in political science and related disciplines, do permit assessment of individual-level attentiveness precisely because, like FMCs, there is a correct response. However, unlike FMCs, IMCs are unrelated to the experiment. This characteristic of IMCs limits the degree to which researchers can use IMC responses to determine respondents' attentiveness to *information in the experiment*. Relatedly, because IMC responses will be identical across experimental conditions in expectation, IMCs are not useful for exploring whether respondents in one condition answered significantly differently from subjects in other conditions, as can be the case with FMC-TR responses. Table 1 provides an overview of these three types of MCs, as well as their differences on key dimensions.

[Table 1 about here]

The kind of information that FMC responses provide to researchers can therefore be helpful when analyzing experimental results, regardless of whether results are statistically significant or not. For example, if 10% of individuals in the experiment answered an FMC correctly, then null results are likely less a reflection upon an author's theory, and more an indication that the treatment stimulus was relatively imperceptible and/or that the sample was highly inattentive. Conversely, if most respondents answer the FMC correctly, yet no significant treatment effect is found, it would suggest a deficiency with the theory and/or an inability of the manipulation to affect the independent variable of interest. In short, FMCs can help reveal *why* an experiment yielded null results.

Alternatively, one might find a significant treatment effect, yet a modest passage rate for the FMC in a particular condition. This information, especially in conjunction with significantly different (between-condition) FMC-TR or SMC responses, could lend additional credence to the findings.[5] That is, the modest FMC passage rate result would imply that the treatment was strong enough to exert an effect *despite* a sizable presence of inattentive respondents. In this way, FMCs enable researchers to more thoroughly adjudicate between theory-related and attentiveness-related interpretations of their experimental findings.

Lastly, FMCs can be useful when *designing* experimental studies. For example, researchers can assess the perceptibility of alternative versions of a treatment by employing an FMC-TR and examining variation in passage rates (i.e., if the FMC-TR is constant across

[5] While our discussion separates the various types, some researchers might opt to employ more than one MC within the same design (e.g., asking an FMC and an SMC). Additionally, many researchers implement SMCs prior to fielding the actual experiment—in the form of pretests—to help ensure that treatments are efficacious (for the importance of pretests, see Mutz 2011, 86 or Chong and Junn 2011, 329).

experimental conditions, then higher passage rates would indicate a more perceptible treatment). Thus, as opposed to IMCs, FMCs afford researchers an ability to better design their treatments at the outset.

FMCs: WHERE SHOULD THEY BE PLACED?

If a researcher decides to include an FMC in an experiment, there remains the practical question of where to place it, particularly with respect to the outcome measure. The purpose of the following sections is to assess, using FMCs, the merits of various claims regarding the placement of MCs in general. Again, on this point, there is surprisingly little empirical guidance. For example, Mutz (2011, 85) states that, "In some cases, manipulation checks are inserted directly after a treatment, but more often they are included after measuring the dependent variable, so that the researcher does not call undue attention to the stimulus or create a demand for certain kinds of responses." In a similar fashion, Berinsky, Margolis, and Sances (2014, 744) state that, "manipulation checks run the risk of priming respondents about the treatment they just experienced, in effect treating them for a second time." More broadly, these authors assert that the inclusion of a manipulation check after a treatment but before the outcome measure serves as an additional event that can distort the treatment effect via distracting respondents from the treatment stimulus.

The central idea in these claims is that placing an MC after a treatment—but *before* the outcome measure—will likely generate treatment effects that are different from what would have been observed had the MC not been placed there. This possibility leads to our first research question:

> **R1:** *Do treatment effects differ significantly depending on the location of an FMC relative to the outcome measure?*

On the other hand, placing an FMC *after* the outcome may also have drawbacks. Most importantly, respondent forgetfulness and/or fatigue may result in a smaller proportion of the sample answering the FMC correctly, despite having actually been attentive to the treatment. In other words, one concern is that an FMC placed after the outcome measure may fail to accurately measure the attentiveness during the preceding experimental portion of the survey (e.g., Hauser and Schwarz 2015, 5). If so, the decision to place an FMC after the outcome measure would likely *underestimate* the share of respondents who were treated, thus undermining the goal of assessing respondent attentiveness. Our second research question is therefore as follows:

> **R2:** *Does placing an FMC after (versus before) the outcome measure result in a significantly smaller proportion of correct responses?*

In the following section, we explore the whether using FMCs has deleterious consequences for treatment effect estimation, as well as whether the positioning of FMCs alters their effectiveness (as measured by passage rates). In addition, we also contrast FMCs with IMCs at various points, depending upon the nature and extent of the comparisons permitted by the designs. Both types of checks attempt to gauge attentiveness; whether they do so equally and whether they screen the same individuals remains unknown.

EMPIRICAL ANALYSES

In this section, we report findings from four experimental studies. We selected experimental designs that have been published in peer-reviewed journals and that found significant treatment effects. However, in addition to including the original experimental conditions, we also constructed FMCs and inserted them into the experimental design. We manipulated both the type of FMC (i.e., treatment-relevant or treatment-irrelevant) and their placement (i.e., before or after the outcome measure). Table 2 lists each study and its key

features. This section serves two purposes: 1) to demonstrate *how treatment-relevant (TR) and treatment-irrelevant (TI) factual MCs can be constructed* for particular studies, and 2) to empirically address R1 and R2. Overall, we find little evidence that placing FMCs before the outcome measure distorts treatment effects.[6] Additionally, we find no evidence that respondents are less likely to answer an FMC correctly when it appears immediately after an outcome versus directly before it.

[Table 2 About Here]

Study 1: Student Loan Forgiveness Experiment

In our first study, we replicated a framing experiment featured in Mullinix et al. (2015). The original experiment featured two conditions regarding a student loan forgiveness program (N=292 for their student sample). In the control condition, respondents were given information about the amount of student loan debt in the U.S and were told that proposals exist for a student loan forgiveness program. In the treatment condition, respondents were given this same information, but were also informed that this program would negatively affect the economy and that it is the student's responsibility to pay back these loans. The authors find that the treatment condition significantly lowered support for the loan forgiveness proposal (measured on a seven-point scale, ranging from "strongly oppose" to "strongly support").

We conducted this same experiment in a university lab with 262 undergraduate students in 2015. Beyond the original experimental manipulation, we also constructed an FMC-TR and FMC-TI, and manipulated whether one of these appeared before or after the outcome measure. The FMC-TR asked respondents, "According to the paragraph you just read, what is a potential consequence of the student loan forgiveness program?" The response options were in multiple-

[6] We provide experiment wordings, variable measurement details, and regression output for the following four studies in the Supplemental Appendix.

choice format; thus, respondents had to choose one of four possible answers, only one of which was factually correct. The FMC-TI used in this experiment asked respondents to identify the federal agency (from a list of four agencies) that was referenced in the article.[7]

This design generated four separate experimental conditions, and enabled us to investigate both R1 and R2. The left facet of Figure 1 displays the main results of our study. By comparing the mean of the control group (5.29) to the mean of the treatment group (4.55, represented by the vertical reference line), the figure indicates that, as in the original experiment, the treatment did indeed lower support for the student loan forgiveness program. Subtracting 4.55 from 5.29 yields a treatment effect equal to (negative) .74 ($p<.05$).[8]

[Figure 1 about here]

However, there is no evidence that placing a FMC before the outcome measure significantly distorted this treatment effect. The mean of the "Treat: TR" group (4.54) is nearly identical to the mean of the aforementioned treatment group (difference=-.01; p=.97). Similarly, by comparing the mean of the "Treat: TI" group (4.63) to the reference line, we can see that when an FMC-TI appeared pre-outcome, the change from the control group (and thus the treatment effect) is slightly smaller than the original treatment effect (-.66 vs. -.74, respectively). But this difference is small and not statistically significant (p=.80). Thus, regardless of whether the experiment used a pre-outcome FMC or not, the results would have been similar. The concerns underlying R1, therefore, do not find empirical support here.

[7] To economize on the number of experimental conditions in our study, yet still address R1 and R2, the FMC-TI was inserted after the outcome measure for those assigned to the original control condition; for those assigned to the original treatment condition, the FMC-TR was inserted after the outcome measure.

[8] Estimates of group-mean differences, and calculations of p-values, for each study are from regression models unless otherwise indicated (see Supplemental Appendix model output).

Nevertheless, to avoid the risk of distorting treatment effects, we could have simply placed the FMC *immediately after* the outcome. However, as noted above (R2), a concern is that doing so might result in a smaller share of respondents answering the FMC correctly (e.g., due to forgetfulness, fatigue, or some other element beyond inattentiveness). Because we varied whether a FMC appeared immediately before or immediately after the outcome, we are able to investigate this question. As shown in the right facet of Figure 1, for the respondents assigned to answer an FMC-TR, 86% answered correctly when it appeared pre-outcome, while only 82% answered correctly when it appeared after the outcome. However, this difference is substantively small and not statistically significant (p=.52).[9] For the respondents who answered an FMC-TI, 57% answered correctly when the MC appeared pre-outcome, while 60% answered correctly when the MC appeared after the outcome question (p=.72). These patterns do not support the concerns underlying R2. Overall, then, passage rates were relatively unaffected by whether an FMC appeared immediately before or after the outcome measure. Additionally, the right facet reveals that between-condition variability in the proportion answering correctly was minimal (range =.04 for FMC-TR, and .03 for FMC-TI).

Study 2: KKK Tolerance Experiment

In our second study we replicate an experiment by Nelson, Clawson, and Oxley (1997), which examined how varying the discussion of a Ku Klux Klan (KKK) rally affects public support for a demonstration. In the original study, the authors found that casting the rally as (1) a potential threat to public order, versus (2) an exercise in free speech, significantly decreased support for the KKK to demonstrate.

[9] All significance tests for passage rates are difference-in-proportions tests unless otherwise indicated.

We fielded our experiment on Amazon.com's Mechanical Turk (MTurk) in 2015. Overall, 536 respondents were randomly assigned to one of six conditions, each of which presented respondents with a (mock) news article about an upcoming KKK rally at Ohio State University (the news article mimicked the protocols of the original experiment). In addition to showing respondents a *free speech* (FS) frame and a *public order* (PO) frame, we also varied whether an FMC appeared before the outcome measure.[10] This factual manipulation check was either an FMC-TR *or* an FMC-TI, and was open-ended in both cases. The FMC-TR asked respondents to indicate what the professor (who was quoted at the end of the article) said about the KKK demonstration. In the FS condition, the professor emphasized the KKK's right to speak, whereas in the PO condition, the professor emphasized that violence could result from the rally and that safety must be a top priority. Thus, the FMC-TR could be answered correctly regardless of having seen the FS or PO frame, and, to answer the FMC-TR correctly, respondents needed to directly reflect upon the key element being manipulated by the researcher: whether the rally was being framed in terms of free speech or public order.

Respondents assigned to a condition with an FMC-TI, on the other hand, were asked, "At which University is the KKK planning on demonstrating?" Again, this information was provided in each condition, regardless of whether it was the FS or PO frame (i.e., the correct answer did *not* vary across conditions). Thus, the FMC-TI captured whether individual respondents were attentive to the experimental material, but did not explicitly invoke the independent variable being manipulated. The open-ended responses to the FMC-TI and FMC-TR were reviewed and hand-coded as either incorrect or correct.

[10] As in the original study, the outcome was a seven-point scale measuring support for the KKK to rally, ranging from "strongly oppose" to "strongly support."

The left facet of Figure 2 displays the key results of the study. As in the original experiment, the results show that the PO frame lowered support for allowing the KKK to demonstrate relative to the FS frame. This can be seen by comparing the mean in the free speech condition ("FS"; 3.61) against the mean in the public order condition ("PO"; 2.75, again represented by the vertical line). The difference between these two means, and therefore the treatment effect, is equal to (negative) .86 ($p < .01$).

Did placement of a factual MC *before* the outcome measure significantly affect this treatment effect? We can investigate this question by looking at estimates both *within* and *between* the FS and PO frames. First, with respect to the three FS means (see the top portion of Figure 2 [left facet]), we see that each 95% confidence interval overlaps with each point estimate, indicating that those assigned to these FS conditions did not significantly differ on the outcome variable from one another.[11] Similarly, as indicated by their 95% confidence intervals, the means in the "PO-TR" and "PO-TI" rows (2.67 and 3.03, respectively) do not significantly differ from the mean in the PO condition. Thus, *within* the FS condition, and *within* the PO condition, including a pre-outcome FMC did not significantly alter group means on the dependent variable.

However, the figure also reveals that, had the original experimental design included a FMC-TR appearing *before* the outcome, the estimated treatment effect would be substantially smaller. This can be seen by comparing the mean in the "FS-TR" row (3.21) against the mean in

[11] We assess significance in this way (i.e., a confidence interval overlapping with the *point estimate* for the original treatment) both to simplify the discussion and because, while non-overlapping confidence intervals indicate statistical significance (at $\alpha=.05$), overlapping confidence intervals are not necessarily indicative of non-significance (Bolsen and Thornton 2014). Again, regression analyses, unless otherwise indicated, were used to conduct formal tests of significance.

the “PO-TR” row (2.67), yielding a difference of only .54 (p=.05) rather than the original difference of .86.

Similarly, comparing the mean in the “FS-TI” row (3.37) against the mean in the “PO-TI” row (3.03) yields a difference of only .34. Thus, had the original experiment contained a FMC-TI appearing before the outcome measure, the estimated treatment effect would again be substantially smaller than the original treatment effect. In fact, with such a design, we would have failed to reject the null hypothesis (p=.24).

[Figure 2 about here]

Thus, *within* a given frame, the placement of the FMC did not significantly alter mean support for the KKK demonstration. However, in comparing *across* frames, we find that placing an FMC before the outcome measure appears to have slightly attenuated the treatment effect in this experiment and—in the case of a pre-outcome FMC-TI—to the point where the null hypothesis can no longer be rejected. Study 2 thus provides some slight evidence in the affirmative regarding R1. However, because the results show an *attenuation* of treatment effects, the findings of Study 2 provide no support for the concern that pre-outcome MCs will *augment* treatment effects via functioning as a second treatment (Berinsky, Margolis, and Sances 2014, 744).

The right facet of Figure 2 displays the passage rates for MCs in each of the experimental conditions. Illustrative of the kind of group-level analyses that can be done with a FMC-TR, this facet confirms that respondents in the FS-TR condition tended to (correctly) mention free speech in answering the FMC-TR, while respondents in the PO-TR condition tended to (correctly) mention public order/safety.

In addition, approximately one third of respondents (N=207) in Study 2 were randomly selected to answer a pre-treatment IMC.[12] The right facet thus indicates that, among respondents who were selected to answer the IMC, the share correctly answering both the FMC and IMC is noticeably lower than the share that correctly answered the IMC. This share is also lower than the share that correctly answered the FMC alone. Indeed, the correlation between the IMC and FMC-TR is a modest .34 ($p<.01$), and equals .18 ($p=.14$) for the IMC and FMC-TI. Overall, roughly 15% who answered the IMC correctly answered the FMC incorrectly, while nearly 60% of respondents who answered the IMC incorrectly answered the FMC correctly. This modest relationship suggests that FMCs and IMCs are not interchangeable measures of attentiveness, especially attentiveness to the content in the experiment. Furthermore, between-condition variability in the proportion answering correctly is noticeably greater for IMCs (range=.14) versus FMCs (range=.05 for TR, and .03 for TI).

Study 3: Combatting Disease Experiment

In our third study, we replicated the canonical framing experiment by Tversky and Kahneman (1981). Respondents are randomly assigned to one of two conditions, both of which involve presenting respondents with a hypothetical situation in which an unusual disease will affect up to 600 people. Two programs can be adopted to combat the disease: one that is guaranteed to affect a given number of people (Program A), and one that may help all 600 people or may not help anyone (Program B). Respondents are asked to choose either Program A or Program B; regardless of the condition, Program B is always the riskier option. What varies across the two conditions is whether the choice between Programs A and B involves *gains* or

[12] In studies 2, 3 and 4, the IMC asked about news website preferences (see Berinsky et al. 2014).

losses. To accomplish this, the two programs are described as either *saving lives* (gains) or allowing people to *die* (losses).

Specifically, in the "Saved" condition, respondents are told that, "If Program A is adopted, 200 people will be saved. If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved." However, in the "Die" condition, respondents are told that, "If Program A is adopted, 400 people will die. If Program B is adopted, there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die." In the original study, Tversky and Kahneman found that a significantly larger share of respondents selected the riskier option (Program B) when the framing was in terms of losses ("Die") rather than in terms of gains ("Saved").

We fielded this same experiment on MTurk in 2015. A total of 484 respondents were randomly assigned to one of six conditions. Two of these conditions were identical to the "Saved" and "Die" conditions described above and did not feature an FMC. The other four conditions were either "Saved" or "Die" frames, but also featured an FMC-TI that appeared either before or after the outcome measure. Specifically, the FMC-TI we constructed asked respondents, "According to the excerpt you just read, which program will have a more certain effect on the lives at stake?" The four response options appeared in multiple-choice format. Respondents were given the following response options: "1) Program A [correct answer], 2) Program B, 3) Programs A and B will have an equally certain effect, or 4) Don't Know." To increase statistical power and simplify the presentation, each of the two original conditions were combined with conditions in which the FMC-TI was asked *after* the outcome measure.[13]

[13] For example, the original "Saved" condition (in which no FMC-TI was asked) was combined with the "Saved" condition in which an FMC-TI was asked after the outcome measure. Because

The results of Study 3 are shown in the left facet of Figure 3, which indicates the proportion of each experimental group that selected the riskier option (i.e., Program B). By comparing the proportion shown for the "Saved" group (.33) against the proportion of the "Die" group (.59, again represented by the vertical line), we can first see that those in the "Die" condition were more likely to select the riskier option, consistent with the original study. Subtracting the former proportion from the latter yields a difference—and thus treatment effect—equal to .26, or 26 percentage points ($p<.01$).

What of the conditions featuring a pre-outcome FMC-TI? Again, the FMC featured in this experiment is treatment-irrelevant precisely because it asks the respondents to reflect upon an element that *does not vary* across conditions. If going from the "Saved" frame to the "Die" frame induces risk-seeking behavior (and vice versa), then, in reference to R1, asking respondents to reflect upon the certainty of each program before the outcome measure could perhaps distort the treatment effect.

However, we find no support for this notion in Figure 3 (see left facet). While those in the "Die" condition with a pre-outcome FMC-TI ("Die-TI") had a proportion equal to .67 (and were thus roughly 8 percentage points more likely to select the riskier option than those in the regular "Die" condition), this difference did not attain statistical significance ($p=.23$). Similarly, the proportion found for those in the "Saved-TI" condition was equal to .35. Thus, these respondents were only 2 percentage points more likely to select the riskier option than those in the regular "Saved" condition, and this difference was non-significant ($p=.75$). Finally, we find no significant difference between (1) the difference in proportions for "Saved" versus "Die", and (2) the difference in proportions for "Saved-TI" versus "Die-TI" (absolute difference in

---

the FMC-TI in this latter condition appeared *after* the outcome measure, treatment effects in these two conditions can be safely combined.

proportions equals .06 [p=.17]). We therefore find no evidence in Study 3 that placing a manipulation check before an outcome can significantly alter treatment effects beyond what we would observe had no FMC appeared before the outcome measure (R1).

[Figure 3 about here]

Because the same FMC was asked before *and after* the outcome measure, we can also use Study 3 to investigate R2. The right facet of Figure 3 indicates that, among those in a "Saved" condition, there was a 63% passage rate when the FMC-TI appeared before the outcome measure, and a 68% passage rate when the FMC-TI appeared after the outcome measure (the difference is non-significant, p=.47). Among those in one of the "Die" conditions, there was a 56% passage rate when the FMC-TI appeared pre-outcome, and a 67% passage rate when it appeared after (the difference is non-significant (p=.17). Thus, Study 3 provides no evidence that the passage rate declined when the FMC-TI was placed immediately after (versus immediately before) the outcome measure.

As in the previous study, a subset of respondents was randomly selected to answer a pre-treatment IMC. Figure 3 indicates that passage rates for the IMC were quite high in each condition, and generally significantly higher than the passage rates for the FMC-TI (overall IMC passage rate = 90.67%). The overall pairwise correlation between FMC-TI and IMC passage equals .14 (p=.12), which is again only modestly positive. Moreover, the between-condition variability in the proportion answering correctly is again greater for IMCs (range=.14) versus the FMC-TI (range=.12).

Study 4: Probation Experiment

In our fourth study, we replicated a key part of a persuasion experiment conducted by Mérola and Hitt (2016). The original study used an MTurk sample and featured a control

condition and treatment ("strong message") condition, both of which mention a study that compares probation and imprisonment for felons. The control condition simply informed respondents that the study was recently commissioned, while the treatment condition indicates that, among other details, the study found imprisonment to be "roughly 20 times the cost of a day on probation." The authors find that the treatment condition significantly increased agreement with a statement indicating that probation should be used as an alternative to prison (measured on a seven-point scale, ranging from "strongly disagree" to "strongly agree").

We replicated this experiment on a sample of 2,693 adult U.S. citizens, collected by Survey Sampling International (SSI).[14] The study was fielded in 2017 and the sample was selected to be nationally representative in terms of gender, age, region, race, and income. As with the previous studies, we constructed an FMC-TR and FMC-TI, and varied their placement with respect to the outcome measure. The FMC-TR asked respondents, "Approximately how much more money does it cost to keep an inmate locked up versus on probation?," while the FMC-TI asked, "A moment ago you were informed about a study that was recently commissioned. Which of the following topics does the study deal with?" Each of these FMCs featured seven response options (only one of which was correct), though several of the experimental conditions instead featured open-ended versions of these FMCs.[15]

The left facet of Figure 4 features the results of the study, which are presented as means of the dependent variable for four main experimental groups. A reference line appears at the mean of the treatment condition that did not include a pre-outcome FMC ("Treat," mean=3.80),

[14] In this experiment (and others) there were additional conditions that were not included in this project.

[15] Responses to the open-ended FMCs were coded as correct or incorrect by two independent coders (agreement exceeded 99%). Overall, perhaps due to the large quantity of numeric information in the treatment vignette, respondents were better able to answer the FMC-TI correctly compared to the FMC-TR.

which demonstrates that the treatment group had higher mean support for probation than the control group (control mean = 3.57; difference with treatment = .23 [p<.01]), thereby replicating the direction and significance of Mérola and Hitt's result. However, with respect to R1, it is not the case that including an FMC-TR, or an FMC-TI, before the outcome measure resulted in a significantly different treatment effect. The mean for the "Treat-TR" group is equal to 3.96; the mean for the "Treat-TI" group is equal to 3.92. In both cases, the 95% confidence intervals of these means overlap with the mean support for probation (i.e., the point estimate) found in the "Treat" condition. Thus, the three treatment-group means are similar in magnitude and (as also confirmed by a Wald test) do not significantly differ from one another (p=.34).

[Figure 4 about here]

With respect to R2, we again find no evidence that placing the FMC immediately after (versus before) the outcome measure resulted in a significantly smaller proportion of respondents answering the question correctly. As can be seen in the right facet of Figure 4, for the FMC-TR, the proportion answering correctly in the treatment and "TR" conditions were nearly identical; for the FMC-TI, the proportion answering correctly tended to be *higher* when placed after (versus before) the outcome measure. Study 4, then, also provides no evidence that respondents were less capable of correctly answering the FMC when it was placed after (versus before) the outcome measure.[16]

As with Studies 2 and 3, Study 4 also featured an IMC. In this case, the IMC appeared at the end of the survey and was displayed to all respondents. The right facet of Figure 4 indicates that, within any given condition, there were substantially different passage rates for FMCs and IMCs. Additional analyses revealed that, depending on whether the FMC was TR or TI and

[16] To simplify the presentation of results, Figure 4 features only results for closed-ended FMCs, though similar patterns were found for open-ended FMCs.

open or closed-ended, correlations between the FMCs and the IMC again tended to be only weakly positive (as low as .04, and only as high as .30) and were non-significant in two out of four cases (see Supplemental Appendix for details). For example, excluding respondents who were assigned to the control condition, 60.94% of those who answered the closed-ended FMC-TR correctly answered the IMC incorrectly, while 69.51% of those who answered the IMC correctly answered the FMC-TR incorrectly. Finally, the between-condition variability in the proportion answering correctly is again greater for IMCs (range=.04) than the ("Treat" and "Treat-TR") FMC-TR (range=.01), though not for the FMC-TI (range=.21).

Summary of Empirical Results

Following concerns raised by others regarding the placement of MCs, four published experiments were replicated in an effort to explore how placement of factual MCs—whether treatment-relevant (TR) or treatment-irrelevant (TI)—may, in and of itself, distort treatment effects. The studies involved diverse topics, and are reflective of many experiments currently used in political science. Across these four studies, we found only modest (but inconsistent) evidence that placing an FMC before (versus after) an outcome measure produces significantly different treatment effects, suggesting that manipulation-check placement is largely inconsequential for treatment effects.[17] Nevertheless, particularly given the possibility that the pre-treatment FMC had heterogeneous effects on respondents,[18] the little evidence that we do find suggests that researchers may be best off placing an FMC *immediately after* the outcome

[17] We also ran the analyses of each study among only those who passed MCs—the results did not substantively differ (see Supplemental Appendix).
[18] For example, the pre-outcome FMC could have amplified treatment effects for some, and depressed effects for others.

measure. This is a viable option given the lack of evidence that placing an FMC after the outcome results in a significantly smaller share of respondents answering the FMC correctly.

Finally, as discussed above, IMCs have potential shortcomings relative to FMCs; however, little is known as to whether IMCs and FMCs both measure the latent construct—i.e., attentiveness—in a relatively similar fashion. We featured both FMCs and IMCs in three out of the four studies we replicated and were therefore able to analyze the relationship between FMC passage and IMC passage.[19] These correlations were consistently weak (i.e., at or below .30), and often non-significant. Thus, while FMCs and IMCs both aim to gauge respondent attentiveness, these findings suggest that the two measures should not be viewed as interchangeable—i.e., even if a respondent passes an IMC, doing so is not a reliable indicator that the respondent will be attentive to the material contained in the survey experiment. Similarly, combining all of our replications, the between-condition variability in the proportion answering correctly is, on average, meaningfully greater for IMCs (average range=.11) versus FMCs (average range=.03 for FMC-TRs, and .08 for FMC-TIs). Finding greater variability in correct responses to IMCs is especially noteworthy given that we *varied* the placement of FMCs (but not IMCs) within our studies, and therefore suggests that FMCs measure attentiveness to the experiment more reliably than IMCs.

DISCUSSION & CONCLUSION

Manipulation checks (MCs) are useful tools for experimentalists. However, usage of MCs in the political science literature remains both rare and under-studied. Researchers who do

[19] In another study we replicated (Gross [2008]; see Supplemental Appendix), we again found only modest correlations between IMC and 1) FMC-TR (r = .22, p<.01), and 2) FMC-TI (r = .15, p<.05) passage rates.

use MCs conceptualize and employ them in a variety of ways, which raises practical questions regarding best practices for MC implementation.

Presently, instructional manipulation checks are a common method for assessing attentiveness in experiments. However, IMCs possess notable limitations compared to FMCs. Empirically, we find the correlation between IMC and FMC passage rates to be quite modest, and variability in IMC responses to be relatively high. This is important insofar as it suggests that the IMCs and FMCs in our studies were not measuring attentiveness in the same way. Thus, while researchers may wish to include an IMC in their study, we stress the usefulness of also including an FMC because it is explicitly designed to measure attentiveness to the *experimental portion* of the study. Further, while SMCs can determine whether a significant difference exists between the treatment and control groups on the latent variables of interest, they provide no direct information regarding the extent to which the sample was attentive. Thus, by using FMCs, researchers have greater ability to analyze and diagnose their experimental findings.

To decide whether to use an FMC-TR or FMC-TI, the researcher need only determine whether a single (factual) question be asked of each group in the experiment that, if answered correctly, would reveal attention to the manipulation. If so, then an FMC-TR can be used. When it is not possible to ask a single factual question regarding manipulated information, an FMC-TI can be used. Again, here the researcher need only identify some element of the information that does *not* vary across experimental conditions and devise a single factual question pertaining to it. To ensure that the FMC responses are as informative as possible, researchers should strive to make the FMC question relatively simple to answer for respondents who attended to the experimental information, but challenging to answer for respondents who did not attend to the

experimental information. In terms of the construction of response options, answers to FMCs can be open-ended or closed-ended.[20]

Regarding placement, the key results of our studies indicate that FMCs can be placed immediately before or after the outcome measure with relatively little consequence for treatment effects (R1), though placing FMCs immediately *after* the outcome measure is likely optimal given that we find no appreciable differences in passage rates (R2).

In sum, relative to other MCs, we suggest that a factual MC placed after the outcome measure is uniquely suited for accomplishing the twin goals of 1) capturing individual-level attentiveness to *the experiment itself*, and 2) avoiding distortion of treatment effects. Importantly, factual MCs require only one additional question to be asked of respondents. Additionally, they may be used in conjunction with other MCs and common methods for ensuring that treatments will be received by respondents, such as pretests and warnings (Clifford and Jerit 2015). Using the protocols we have outlined for constructing and implementing FMCs, researchers can better assess attentiveness to their experiments and therefore more thoroughly diagnose findings, test hypotheses, and advance theory.

[20] In the experiments we conducted, open-ended FMCs necessarily required some form of content analysis, while closed-ended (multiple choice) FMCs had randomized response options.

REFERENCES

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

Anduiza, Eva, and Carol Galais. 2016. “Answering Without Reading: IMCs and Strong Satisficing in Online Surveys.” *International Journal of Public Opinion Research,* 29 (3): 497-519.

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. “Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys.” *American Journal of Political Science* 58 (3): 739–53.

Bolsen, Toby, and Judd R. Thornton. 2014. “Overlapping Confidence Intervals and Null Hypothesis Testing.” *Newsletter of the APSA Experimental Section* 4 (1): 12–16.

Broockman, David E., and Daniel M. Butler. 2017. “The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication.” *American Journal of Political Science* 61 (1): 208–21.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci. 2014. “Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers.” *Behavior Research Methods* 46 (1): 112–30.

Chandler, Jesse, and Danielle Shapiro. 2016. “Conducting Clinical Research Using Crowdsourced Convenience Samples.” *Annual Review of Clinical Psychology* 12 (1): 53–81.

Chong, Dennis, and Jane Junn. 2011. "Politics from the Perspective of Minority Populations." In *Cambridge Handbook of Experimental Political Science*, eds. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. New York: Cambridge University Press, 320-35.

Clifford, Scott, and Jennifer Jerit. 2014. “Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies.” *Journal of Experimental Political Science* 1 (2): 120–31.

———. 2015. “Do Attempts to Improve Respondent Attention Increase Social Desirability Bias?” *Public Opinion Quarterly* 79 (3): 790–802.

Clifford, Scott, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. 2015. “Moral Foundations Vignettes: A Standardized Stimulus Database of Scenarios Based on Moral Foundations Theory.” *Behavior Research Methods* 47 (4): 1178–98.

Crump, Matthew J. C., John V. McDonnell, and Todd M. Gureckis. 2013. "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research." *PLOS ONE* 8 (3): e57410.

Curran, Paul G. 2016. "Methods for the Detection of Carelessly Invalid Responses in Survey Data." *Journal of Experimental Social Psychology* 66 (Sept.): 4-19.

Dietrich, Simone, and Matthew S. Winters. 2015. "Foreign Aid and Government Legitimacy." *Journal of Experimental Political Science* 2 (02): 164–171.

Gross, Kimberly. 2008. "Framing Persuasive Appeals: Episodic and Thematic Framing, Emotional Response, and Policy Opinion." *Political Psychology* 29 (2): 169–92.

Hauser, David J., and Norbert Schwarz. 2015. "It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on 'Tricky' Tasks." *SAGE Open* 5 (1): 1–6.

Hauser, Daniel, and Norbert Schwarz. 2016. "Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than Do Subject Pool Participants." *Behavioral Research* 48: 400-7.

Healy, Andrew, and Gabriel S. Lenz. 2014. "Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy." *American Journal of Political Science* 58 (1): 31–47.

Horowitz, Michael C., and Matthew S. Levendusky. 2011. "Drafting Support for War: Conscription and Mass Support for Warfare." *The Journal of Politics* 73 (02): 524–34.

Krosnick, Jon A., Sowmya Narayan, and Wendy Smith. 1996. "Satisficing in Surveys: Initial Evidence." *Advances in Survey Research* 1996 (70): 29–44.

Krupnikov, Yanna, and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1 (01): 59–80.

Lovett, Matt, Saleh Bajaba, Myra Lovett, and Marcia J. Simmering. 2018. "Data Quality from Crowdsourced Surveys: A Mixed Method Inquiry into Perceptions of Amazon's Mechanical Turk Masters." *Applied Psychology* 67 (2): 339-66.

Maniaci, Michael R., and Ronald D. Rogge. 2014. "Caring about Carelessness: Participant Inattention and Its Effects on Research." *Journal of Research in Personality* 48: 61–83.

Maoz, Ifat. 2012. "The Face of the Enemy: The Effect of Press-Reported Visual Information Regarding the Facial Features of Opponent Politicians on Support for Peace." *Political Communication* 29 (3): 243–56.

Mérola, Vittorio, and Matthew P. Hitt. 2016. “Numeracy and the Persuasive Effect of Policy Information and Party Cues.” *Public Opinion Quarterly* 80 (2): 554–62.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. Forthcoming/2018. “How Conditioning on Post-Treatment Variables Can Ruin Your Experiment and What to Do about It.” *American Journal of Political Science*.

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

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. “The Generalizability of Survey Experiments.” *Journal of Experimental Political Science* 2 (02): 109–138.

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

Mutz, Diana C., and Robin Pemantle. 2015. “Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods.” *Journal of Experimental Political Science* 2 (02): 192–215.

Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. “Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance.” *The American Political Science Review* 91 (3): 567.

Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko. 2009. “Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power.” *Journal of Experimental Social Psychology* 45 (4): 867–72.

Shadish, William R, Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inferenc*e. Boston: Houghton Mifflin.

Turner, Joel. 2007. “The Messenger Overwhelming the Message: Ideological Cues and Perceptions of Bias in Television News.” *Political Behavior* 29 (4): 441–64.

Tversky, Amos, and Daniel Kahneman. 1981. “The Framing of Decisions and the Psychology of Choice.” *Science* 211 (4481): 453–58.

Table 1. A Typology of Manipulation Checks

Type of MC	Placement in Survey	Entity Measured	Related to Treatment?	Level Used for Analysis
Subjective	After experimental treatment; before or after outcome measure	Perceptions regarding information contained in the experiment	Yes	Group
Instructional	Anywhere in the survey	Response to explicit instructions that are (discreetly) embedded within a survey question	No	Individual
Factual	After experimental treatment; before or after outcome measure	Objective responses to a factual question about information contained in the experiment	Yes (FMC-TR) or No (FMC-TI)	Individual or Group

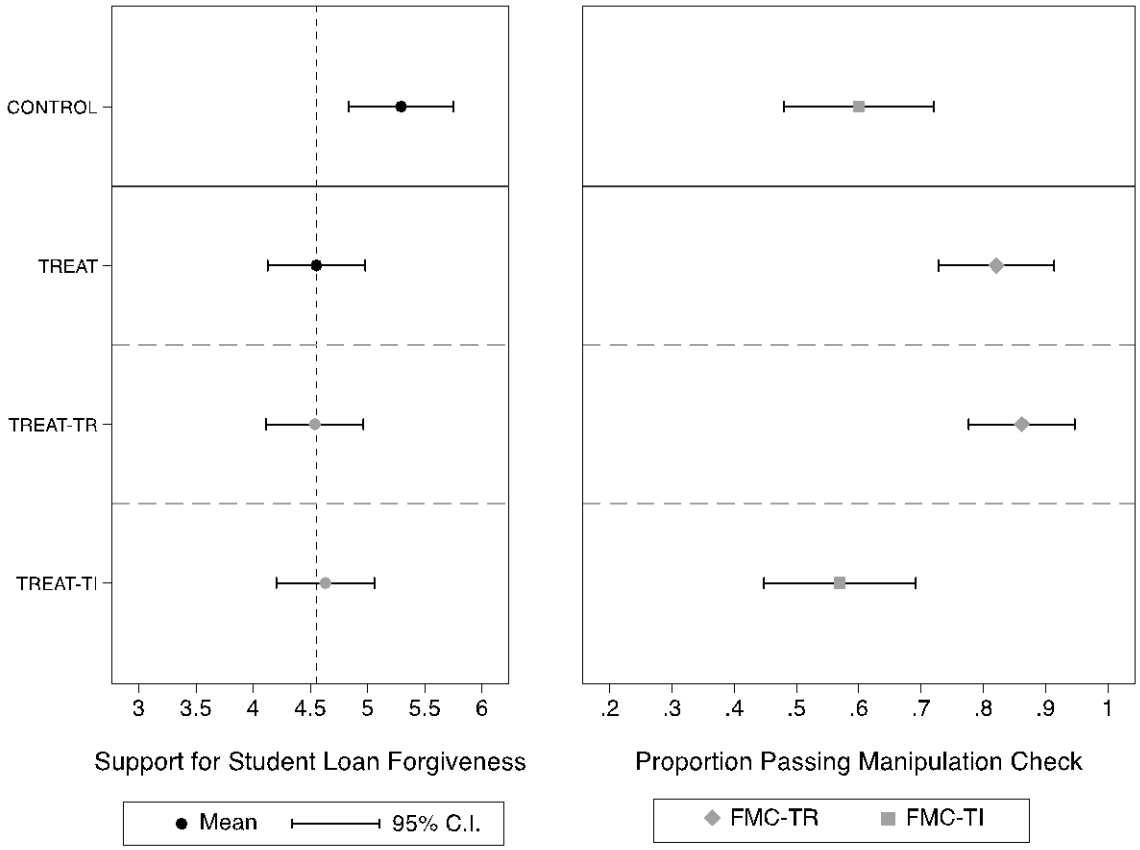
Notes: Table details attributes of three common types of manipulation checks. "FMC-TR" refers to treatment-relevant factual manipulation checks; "FMC-TI" refers to treatment-irrelevant factual manipulation checks. The "Level Used for Analysis" column refers to the level of analysis employed when analyzing MC results: "Individual" indicates individual responses can be analyzed to assess attentiveness; "Group" indicates that researchers can compare experimental groups' average response to the MC.

Table 2. List of Experiments and Features

Study	*Title*	**FMC-TR Featured?**	**FMC-TI Featured?**	**Research Question Addressed**	**Original Study**
1	*Student Loan Forgiveness*	YES	YES	R1 & R2	Mullinix, Leeper, Druckman, and Freese (2015)
2	*KKK Tolerance*	YES	YES	R1	Nelson, Clawson, and Oxley (1997)
3	*Combatting Disease*	NO	YES	R1 & R2	Tversky and Kahneman (1981)
4	*Probation*	YES	YES	R1 & R2	Mérola and Hitt (2016)

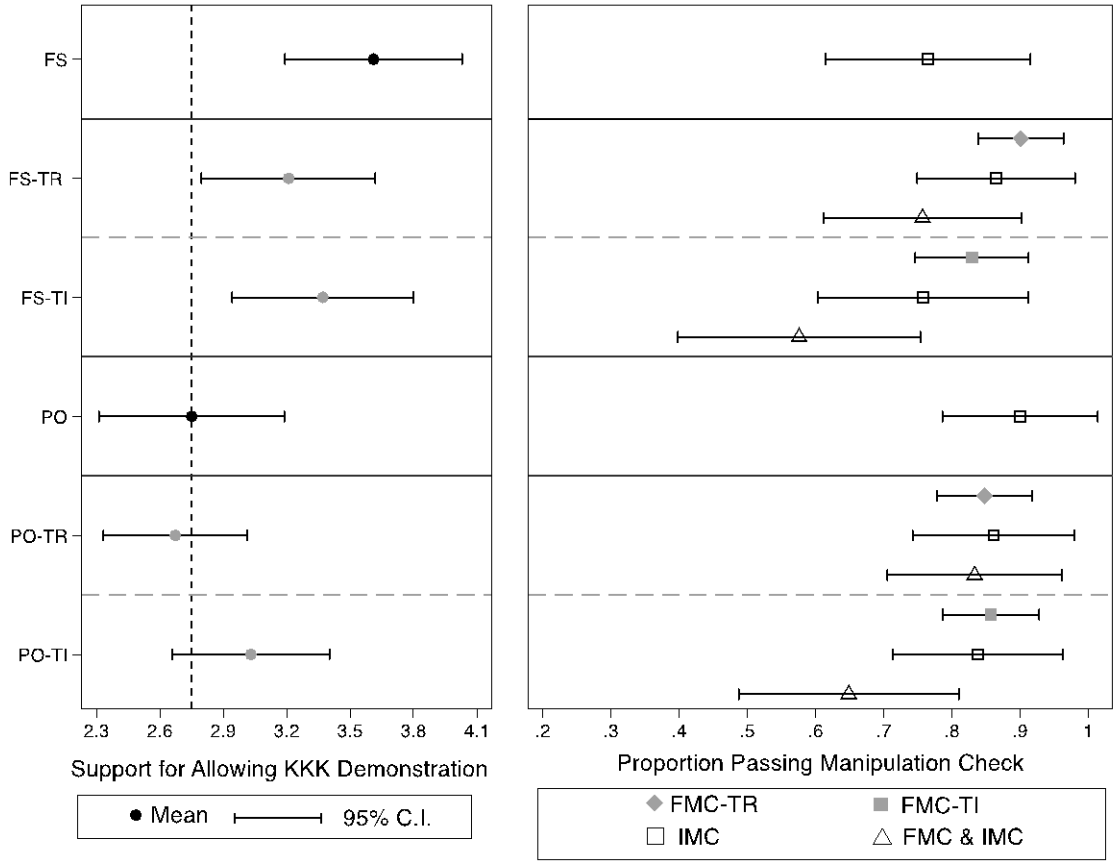
Notes: "FMC-TR" / "FMC-TI" indicate treatment-relevant and treatment-irrelevant factual manipulation checks, respectively. The key manipulation in each study is whether a FMC appears post-treatment and *pre*-outcome vs. post-treatment and *post*-outcome. R1 asks, "Do treatment effects differ significantly depending on the location of a FMC relative to the outcome measure?" R2 asks, "Does placing a FMC after (versus before) the outcome measure result in a significantly smaller proportion of correct responses?"

Figure 1. Study 1 (Student Loan Forgiveness) Showing No Significant Distortion of Treatment Effects with Pre-Outcome Factual Manipulation Checks



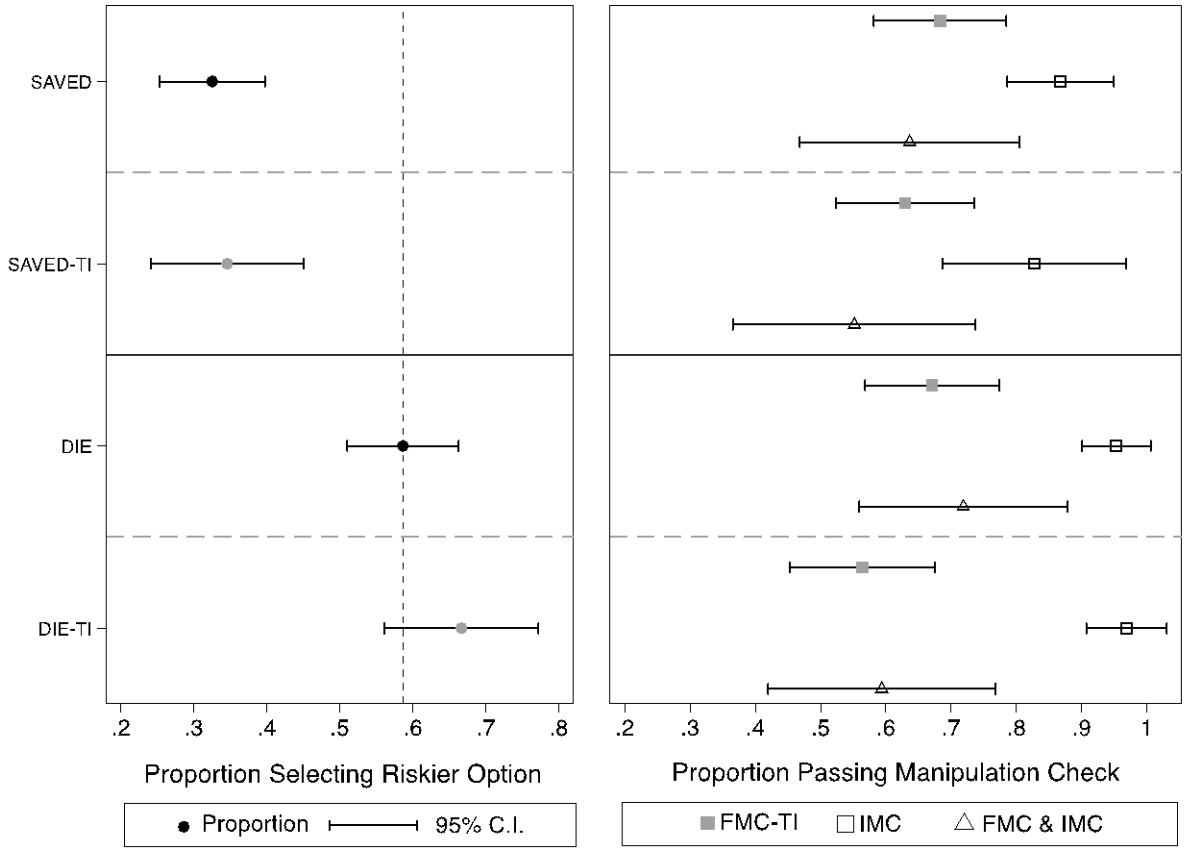
Note: Solid horizontal line separates original treatment and control conditions; dashed horizontal lines separate conditions in which the placement or type of factual manipulation check (FMC) was manipulated. *Left facet*: Dependent variable is a seven-point scale measuring support for the student loan forgiveness program, ranging from "Strongly Oppose" (=1) to "Strongly Support" (=7). "TR" ["TI"] indicates that a treatment-relevant [treatment-irrelevant] factual manipulation check appeared between treatment and the outcome measure. A vertical reference line appears at the mean level of support in the original treatment condition. *Right facet*: This graph indicates the proportion that correctly answered the factual manipulation check (FMC) (either TR or TI) in each experimental condition. 95% confidence intervals shown. Student sample; N=262.

Figure 2. Study 2 (KKK Tolerance) Showing Modest Distortion of Treatment Effects with Pre-Outcome Factual Manipulation Checks



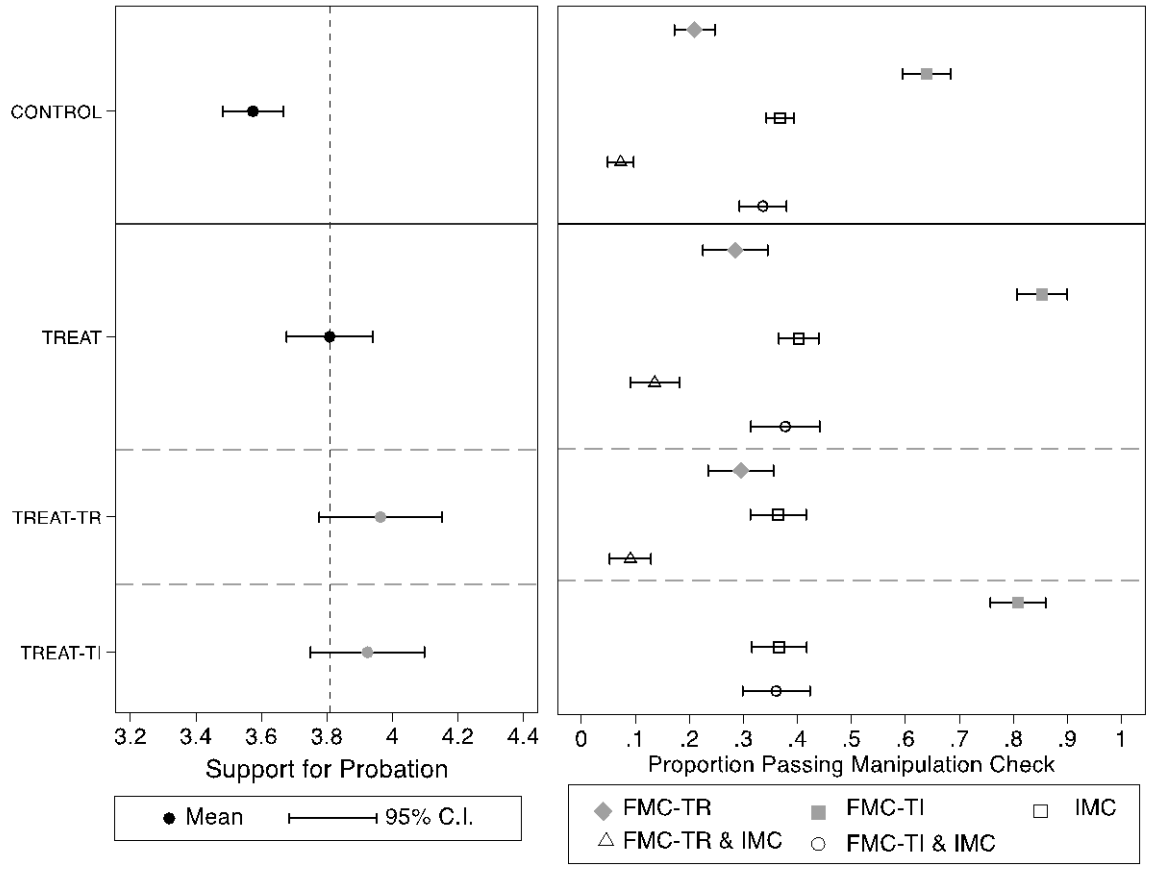
Note: Solid horizontal lines separate original treatment and control conditions; dashed horizontal lines separate conditions in which the presence or type of factual manipulation check (FMC) was manipulated. *Left facet*: Dependent variable is a seven-point scale measuring support for the KKK to demonstrate, ranging from "Strongly Oppose" (=1) to "Strongly Support" (=7). FS=Free Speech frame; PO=Public Order frame. "TR" ["TI"] indicates that a treatment-relevant [treatment-irrelevant] factual manipulation check appeared between treatment and the outcome measure. A vertical reference line appears at the mean level of support in the original treatment condition. *Right facet*: This graph indicates the proportion that correctly answered the FMC (either TR or TI) and/or instructional manipulation check (IMC) in each experimental condition. Only a subset of respondents in each condition were asked to answer an IMC. 95% confidence intervals shown. MTurk sample; N=536.

Figure 3. Study 3 (Combatting Disease) Showing No Significant Distortion of Treatment Effects with Pre-Outcome Factual Manipulation Checks



Note: Solid horizontal line separates original treatment and control conditions; dashed horizontal lines separate conditions in which the placement of a factual manipulation check (FMC) was manipulated. *Left facet*: Dependent variable is a choice between a less risky and a more risky option. X-axis indicates proportion selecting the riskier of these two options; y-axis indicates the experimental condition. "TR" ["TI"] indicates that a treatment-relevant [treatment-irrelevant] FMC appeared between treatment and the outcome measure. A vertical reference line appears at the proportion (selecting the riskier option) in the original treatment condition. *Right facet*: This graph indicates the proportion that correctly answered the FMC and/or instructional manipulation check (IMC) in each experimental condition. Only a subset of respondents in each condition were asked to answer an FMC or an IMC. 95% confidence intervals shown. MTurk sample; N=484.

Figure 4. Study 4 (Probation) Showing No Significant Distortion of Treatment Effects with Pre-Outcome Factual Manipulation Checks



Note: Solid horizontal line separates original treatment and control conditions; dashed horizontal lines separate conditions in which the placement or type of factual manipulation check (FMC) was manipulated. *Left facet*: Dependent variable is a seven-point scale measuring agreement with the proposition that probation should be used as an alternative to imprisonment, ranging from "Strongly Disagree" (=1) to "Strongly Agree" (=7). "TR" ["TI"] indicates that a treatment-relevant [treatment-irrelevant] factual manipulation check appeared between treatment and the outcome measure. A vertical reference line appears at the mean level of agreement in the original treatment condition. *Right facet*: This graph indicates the proportion that correctly answered the FMC (either TR or TI) and/or instructional manipulation check (IMC) in each experimental condition. The IMC appeared for all respondents. 95% confidence intervals shown. SSI sample; N=2,693.