ANALYZE THE ATTENTIVE & BYPASS BIAS: MOCK VIGNETTE CHECKS IN SURVEY EXPERIMENTS

John V. Kane
New York University

Yamil R. Velez
Columbia University

Jason Barabas
Dartmouth College

Word Count: 10,901

ABSTRACT

Respondent inattentiveness threatens to undermine experimental studies. In response, researchers incorporate measures of attentiveness into their analyses, yet often in a way that risks introducing post-treatment bias. We propose a design-based technique—mock vignettes (MVs)—to overcome these interrelated challenges. MVs feature content substantively similar to that of experimental vignettes in political science, and are followed by factual questions (mock vignette checks [MVCs]) that gauge respondents' attentiveness to the MV. Crucially, the same MV is viewed by all respondents prior to the experiment. Across five separate studies, we find that MVC performance is significantly associated with (1) stronger treatment effects, and 2) other common measures of attentiveness. Researchers can therefore use MVC performance to re-estimate treatment effects, allowing for hypothesis tests that are more robust to respondent inattentiveness and yet also safeguarded against post-treatment bias. Lastly, our study offers researchers a set of empirically-validated MVs for their own experiments.

Researchers are increasingly turning to online samples to conduct studies. Naturally, greater reliance upon such data has raised concerns that some share of respondents will not be fully attentive to the survey, perhaps rushing through and not effortfully considering the presented information (Alvarez et al. 2019; Hauser and Schwarz 2016; Thomas and Clifford 2017). For those conducting survey experiments, this issue presents a critical challenge: to the extent that a sample is inattentive, "treatments" will not actually be received, and estimates of treatment effects will, in expectation, likely be biased toward zero.[1] Inattentiveness therefore threatens to fundamentally undermine what researchers can learn from their studies.

Given the seriousness of this challenge, researchers have begun developing ways of assessing attentiveness in online surveys. Kane and Barabas (2019) recommend using factual manipulation checks (FMCs) after experiments' outcome measures; others have implemented question timers to gauge how long respondents spend on a given survey item (Niessen, Meijer, and Tendeiro 2016; Wood et al. 2017); still others have employed instructional manipulation checks (IMCs), also known as "screeners" (Berinsky, Margolis, and Sances 2014; Oppenheimer, Meyvis, and Davidenko 2009). The primary function of such tools is to enable the researcher to diagnose which respondents were attentive to an experiment. But, once this individual-level attentiveness is measured, researchers often use such tools for a second purpose: to re-estimate treatment effects on those deemed to be attentive.

Yet recent research has raised serious concerns with this latter practice. Specifically, using a post-treatment variable (such as a manipulation check or question timer) to re-estimate treatment effects—e.g., by removing those respondents deemed to be inattentive to the experiment, or by

[1] See, for example, Gerber and Green (2012), who illustrate how intention-to-treat (ITT) effects are smaller to the extent that subjects do not comply with treatment despite being assigned to treatment.

interacting the treatment with the attentiveness measure—can introduce covariate imbalances between the randomized treatment and control groups, therein biasing one's estimated treatment effect (Aronow, Baron, and Pinson 2019; Coppock 2019; Montgomery, Nyhan, and Torres 2018).

How, then, can researchers (1) measure individual-level attentiveness, and (2) use this measure to estimate treatment effects among those deemed to be attentive, yet also (3) avoid introducing post-treatment bias? In this study, we propose a new, design-based technique—*mock vignettes*—for simultaneously accomplishing these three objectives. A mock vignette (MV) contains largely descriptive information that, in terms of content, is substantively similar to the types of content found in political science experiment vignettes. Crucially, the MV appears *before* the researcher's experiment, and all respondents read the same MV. Next, respondents answer factual questions about the vignette that check for comprehension of the MV. We refer to these items as *mock vignette checks* (MVCs). From the respondent's perspective, therefore, this technique simulates the format of a typical survey experiment: respondents are asked to read and consider a short paragraph of information (i.e., a vignette) generally related to current and/or governmental affairs, and then, on a separate screen, are asked follow-up questions (e.g., Mutz 2011, Ch. 4; Steiner, Atzmüller, and Su 2016).

Using responses to MVCs, the researcher can construct an individual-level measure of attentiveness as it exists immediately prior to the actual experiment. Moreover, this measure can then be used to analyze respondents who "pass" the MVC—specifically, to present not only the average treatment effect (ATE) estimated for the sample as a whole (or, more accurately, the intention-to-treat (ITT) estimate), but also the treatment effect among respondents deemed to be attentive (as measured by the MVC). To the extent that inattention is downwardly biasing a treatment effect, then, the researcher should observe a *stronger* effect when analyzing those who pass (versus fail) the MVC. Most importantly, because the MV and MVC appeared *prior to* randomization in the researcher's

experiment, utilizing mock vignettes bypasses the problem of post-treatment bias (see Montgomery, Nyhan, and Torres 2018, 771).

To test the merits of our *mock vignette* approach, we replicate a series of published experiments using samples from a variety of online respondent pools (MTurk, Qualtrics, NORC, and Lucid). In each study, we feature one MV and at least one MVC immediately prior to the experiment. We consistently find, first, that treatment effects are significantly stronger among those who performed well (versus poorly) on the MVCs. Second, we find that MVC passage is strongly predictive of performance on other established measures of attentiveness, including timers on various items in the experiment (e.g., the MV itself, experimental vignettes and experimental outcome question(s)) and FMCs. Third, we investigate the possibility that MVs may inadvertently prime various respondents, or generate additional fatigue, and thus substantially alter the ATE relative to what would have been observed had no MV been employed. Across each of our studies that randomly assigned whether a MV was featured, we find no evidence for this concern. Lastly, we investigate whether there are demographic and/or political correlates of MVC performance. Overall, and consistent with extant research, we find only a couple of demographic variables to be consistently associated with MVC performance (namely, age and race), though these correlations were substantively modest in size. However, we do not find any consistent evidence of political variables (i.e., party identification, ideological self-placement, and political interest) being associated with MVC performance.

In sum, mock vignettes offer a simple, design-based method to help researchers test hypotheses on respondents more likely to have been attentive to their experiment.[2] Further, MVs can be used in

[2] This proposition echoes a suggestion by Mutz (2011, 153) who, after differentiating treatment effects among those who would *actually*, versus only *potentially*, be exposed in the real world,

conjunction with other tools (such as manipulation checks) and techniques (such as pre-treatment warnings (Clifford and Jerit 2015)) aimed at measuring and augmenting respondent attentiveness to the experiment. Finally, in Supplemental Appendix C, we offer researchers a variety of ready-made MVs and MVCs, each validated with online-sample data and complete with various descriptive analyses, including passage rates, correlations with demographic variables, measures of complexity, and item response theory (IRT) analyses. By using a mock vignette before their experiment, researchers are better equipped to overcome the problem of respondent inattentiveness, and therefore perform fairer, more reliable, and more robust tests of their hypotheses.

NONCOMPLIANCE & POST-TREATMENT BIAS

Whether researchers attempt to measure it or not, experiments fielded online will likely contain a sizable share of inattentive respondents. Respondents may, for example, be distracted during the experiment (Clifford and Jerit 2014), or simply "satisfice" as a means of completing the survey as quickly as possible to receive payment (Anduiza and Galais 2016; Krosnick, Narayan, and Smith 1996). Such inattentiveness represents a form of experimental noncompliance, which, as Harden, Sokhey and Runge (2019, 201) contend, "poses real threats to securing causal inferences and drawing meaningful substantive conclusions." This is largely because, if simply ignored by the researcher, respondent inattentiveness to the experiment and/or its outcome measures threatens to bias treatment effects downward toward zero, thereby increasing the probability of a Type II error. Imprecise

writes, "if it is feasible to include pre-test survey questions to provide some purchase on who is most likely to be exposed to a given treatment in the real world, then the researcher can essentially have her cake and eat it too."

estimates, and/or null or weak effects, then, may be mistakenly interpreted as a flawed theory or design, rather than as a consequence of respondent noncompliance. Thus, even if a researcher obtains a large, probability sample, inattentiveness serves to undermine hypothesis tests, which (virtually always) implicitly assume respondent attentiveness.

Strategies for improving precision include developing stronger treatments via pretesting, blocking, including pre-treatment covariates that predict the outcome, or simply increasing sample size (e.g., see Shadish, Cook, and Campbell 2002). Yet these options are not always feasible, nor do they actually address the problem of noncompliance downwardly biasing effect sizes. A larger sample, for example, may help yield a treatment effect that is "statistically significant," but the magnitude of that effect will nevertheless likely be smaller than it would have been had the sample been fully attentive.

Recent literature has promoted the use of various tools for directly measuring respondent attentiveness. Kane and Barabas (2019), for example, recommend post-outcome factual manipulation checks (FMCs), which are objective questions about the experimental information given to respondents. Others have utilized instructional manipulation checks (IMCs), also known as "screeners", which are questions that discreetly ask respondents to answer a seemingly banal survey question in a specific fashion (Berinsky, Margolis, and Sances 2014; Oppenheimer, Meyvis, and Davidenko 2009). In these studies, answering the manipulation check questions correctly is indicative of greater attentiveness to the experiment, whereas answering incorrectly signals lower attentiveness. A third approach involves the use of question timers, wherein the amount of time that respondents spend on a given screen is recorded. Such times (or, latencies) are treated as a measure of attentiveness, particularly insofar as low latency signals insufficient attention (Harden, Sokhey, and Runge 2019, 3; Niessen, Meijer, and Tendeiro 2016; Wood et al. 2017; Zwaan et al. 2018).

What can be done with these measures? On one hand, such individual-level measures of attentiveness can be used to gauge the overall share of attentive respondents in any given experimental condition (or for the study as a whole). This serves as a useful diagnostic tool to help adjudicate between competing interpretations of a given result (e.g., an empirically unsupported hypothesis versus a small effect that arose from substantial respondent inattentiveness). FMCs also have the added benefit of being able to ensure that responses to a given question correlate with treatment assignment, thereby functioning not only as a measure of general attention to the content but also as evidence that the manipulation itself was sufficiently perceived.

However, beyond merely identifying inattentive respondents, researchers have also tended to use such measures in estimating treatment effects. For example, some researchers simply subset the data on this measure, in effect excluding from the analysis respondents deemed insufficiently attentive (see Aronow, Baron, and Pinson 2019). Similarly, researchers attempt to specify such measures as control variables in regression models, or interact these measures with the treatment indicator variable to test whether the treatment effect differs across levels of attentiveness. The problem with such techniques is that they, in effect, threaten to "de-randomize" the experimental groups (Coppock 2019). That is, conditioning on a post-treatment variable threatens to create treatment and control groups that are compositionally *dissimilar*, resulting in a biased estimate of the treatment effect (Acharya, Blackwell, and Sen 2016). Worse still, researchers have limited statistical ability to completely rule out the possibility of post-treatment bias (Montgomery, Nyhan, and Torres 2018, 772–73).

Though rarely utilized in survey-experimental research, one proposed statistical approach has been to use a two-stage least squares (2SLS) regression model (or more broadly, instrumental variables (IV) models), wherein treatment assignment serves as an instrument for compliance, which is assessed using a post-treatment measure of attentiveness (e.g., a timer or an FMC). However, it is important to

note that results from such models are more difficult to interpret (Montgomery, Nyhan, and Torres 2018, 771), and properly estimating causal effects among compliers using 2SLS requires strong assumptions that may not be met in practice. For estimates of complier average causal effects (CACEs) to be consistent in this context, the effect of treatment assignment on outcomes must be transmitted entirely via attentiveness (see Green 2013). Moreover, the 2SLS approach implicitly assumes that inattentive respondents are nevertheless sincerely responding to the *outcome* measure(s), which constitutes an untestable (and perhaps implausible) assumption.

The 2SLS approach also presents complexities in terms of actual implementation. For example, if a timer (i.e., latency measure) is used to capture attentiveness, the researcher must decide on the cut-off time that constitutes sufficient attentiveness. Second, for at least one experimental group, actual attentiveness must be disregarded. In other words, in order to ensure that treatment assignment can serve as an instrument for attentiveness, all respondents in one experimental group must be assigned a latency value equal to 0, or be asked a factual manipulation check that they (in expectation) are unable to answer (see Harden, Sokhey, and Runge 2019). This particular requirement can be especially problematic when a researcher utilizes a control condition containing information that should be attended to (e.g., a "placebo" control condition). In effect, these various requirements mean that one can potentially obtain substantially different CACEs depending on (1) the latency cut-off that is decided upon, (2) which experimental group the researcher designates as the group for which attentiveness will equal 0, and/or (3) whether a latency measure or manipulation check is used to assess attentiveness.[3] Regarding this latter point, proper implementation of the 2SLS method becomes even

[3] See Supplemental Appendix H for an empirical demonstration of this point, featuring examples from our own experiments (described below).

more ambiguous when a researcher wishes to test for significant differences between two *treatment* conditions, as well as in survey experiments with a variety of treatment conditions (e.g., factorial designs and conjoint experiments (Hainmueller, Hopkins, and Yamamoto 2014)).

Given these complexities, we propose a simpler, design-based approach to creating a measure of respondent attentiveness that can be easily incorporated into analyses of survey experiments (including factorial designs and conjoint experiments), and does not threaten to introduce post-treatment bias. We refer to this technique as a *mock vignette* (MV).

MOCK VIGNETTES

Any measure of attentiveness to the experiment itself, as well as any measure of attentiveness occurring after the experiment, is, *ipso facto*, a post-treatment measure. Experimental manipulation checks and timers on experimental content (e.g., timers on vignettes, outcome measures, etc.) are, therefore, post-treatment and risk introducing post-treatment bias when involved in the estimation of treatment effects. Thus, while such a measure is ideal because it directly gauges attentiveness to our experiment's vignettes, a suitable alternative is needed if we wish to re-estimate treatment effects on the attentive respondents.

To do so, we first reason that, because respondent attentiveness varies throughout the course of completing a survey (e.g., Alvarez et al. 2019; Berinsky, Margolis, and Sances 2014), such an alternative measure should be as close in temporal proximity to the experiment as possible—ideally, immediately pre-treatment. Second, we reason that the best alternative to measuring attentiveness to the experimental content itself would be to measure attentiveness to content *of a similar format and general nature*. As detailed below, in the studies we conducted, every respondent viewed several sentences of information related to a current event, and then answered factual, closed-ended questions

about this content, before proceeding to the survey experiment. Designed as such, a respondent's attentiveness to this pre-treatment content can function as a proxy for the respondent's attentiveness to the actual experiment's vignettes and outcome measure(s).

We therefore propose that researchers use a pre-treatment mock vignette (MV) and follow-up "check" questions (MVCs) in their experiments. The MV should, as is typical of experimental vignettes and/or outcome measures in political science (Steiner, Atzmüller, and Su 2016), display information to respondents. The MV's content can, for example, involve descriptive information about some news or policy-related event. In this way, MVs are designed to *simulate* the experience of participating in a typical online survey experiment. Yet the MV should also be free of any explicitly partisan, ideological, or otherwise strongly evocative content as the MV's function is not to, *itself*, exert any discernible treatment effects. Crucially, each respondent sees *the exact same MV*—i.e., the MV, and follow-up MVCs, are identical for all respondents.

Next, respondents are asked at least one MVC, which is a factual question about the content they were just instructed to read in the MV, and which appears on a different screen from the MV.[4] As any given MVC should have only one correct answer, researchers can use responses to the MVC to construct an individual-level measure of attentiveness to the MV (i.e., answering correctly is indicative of greater attentiveness). Should *multiple* MVCs be employed (see examples below) an (additive) attentiveness scale can be constructed. Following the MV and MVC(s), each respondent is then randomly assigned to an experimental condition.

[4] To ensure that respondents could not look up correct answers to MVCs, respondents were not permitted to use the "back" button in any of our studies.

Once this procedure is complete, the researcher is equipped with a pre-treatment measure of respondent attentiveness. More specifically, the researcher will possess what is akin to a pre-treatment proxy measure of the attentiveness the respondent *would have* exhibited during the researcher's experiment. This measure can then be used to re-estimate the ATE among respondents deemed to be attentive by filtering out those who are inattentive. Similarly, the researcher can test the robustness of their ITT estimate by interacting the treatment indicator with MVC performance: if a treatment was indeed efficacious, such an analysis will tend to reveal substantively stronger conditional average treatment effects (CATEs) among those who performed better (versus worse) on the MVC(s).

In contrast to the 2SLS approach noted above, this procedure is implemented in the same manner regardless of how many treatment groups are in the experiment, and regardless of which group is designated as the "treatment" group, and also allows for a *multi-item* measure of attentiveness to be employed in the analysis.[5] Employing mock vignettes in experiments is, therefore, a relatively simple, design-based approach that does not require the statistical assumptions, nor the more complicated modeling choices inherent in other techniques that attempt to address inattentiveness. Most importantly, because they are implemented *prior to* random assignment, MVs bypass the problem of biasing treatment effect estimates with a post-treatment variable (Montgomery, Nyhan, and Torres 2018, 770–71).

It is worth noting that the general logic underlying the MV technique is similar to that of IMCs (also known as "screeners"), but differs in several key respects. First, an IMC is not a vignette—it is a single survey *question*, (ostensibly) about an unrelated topic (e.g., one's favorite color). Perhaps as a

[5] Nevertheless, we wish to highlight that the MV method does not preclude the use of the 2SLS method—i.e., the two techniques are not mutually exclusive.

result, recent research has suggested that online samples have become more savvy in detecting IMCs (Thomas and Clifford 2017), which is plausible given the distinctive appearance and contents of IMCs. Second, IMCs inherently involve a degree of *deception*, whereas MVs do not. On this point, some research has suggested downstream consequences for experimental behavior upon learning that a researcher is attempting to "trap" the respondent with an IMC (Hauser and Schwarz 2015). Third, and most importantly, an MV is explicitly designed to be implemented one time and pre-treatment, whereas IMCs are advised to appear at multiple points throughout a survey (Berinsky, Margolis, and Sances 2014), perhaps even post-treatment. By virtue of their placement, therefore, incorporating IMC performance into one's analysis of treatment effects may inadvertently introduce post-treatment bias (Montgomery, Nyhan, and Torres 2018, 771).

In sum, employing a mock vignette approach potentially offers researchers a new method for both analyzing the attentive *and* bypassing post-treatment bias. As attentiveness is typically a precondition for being able to be treated, it should be the case that performance on a measure of attentiveness to the MV—i.e., the MVC—is associated with stronger treatment effects. We directly investigate this hypothesis in the following section.

DATA & METHODS

In this and the following section, we first provide a general overview of the five studies we conducted, beginning in May of 2019 through February of 2020, using U.S. adults at least 18 years of age. Next, we provide greater detail regarding the mock vignettes and published experiments featured within each study. We then discuss the results of each of these five studies, with particular emphasis on the extent to which better MVC performance is associated with stronger treatment effects as well as better performance on other measures of attentiveness. We then investigate the possibility that

utilizing MVs might systematically distort treatment effects relative to what would have been observed had no MV been featured, as well as findings regarding MV placement and demographic patterns in MVC performance.

Overview of Studies and Designs

Table 1 provides an overview of the first four studies (the fifth is detailed below), including their respective sample sizes. Two of these studies (MTurk 1 and MTurk 2) feature samples from Amazon.com's Mechnical Turk. Another study (Qualtrics) uses a nonprobability sample collected by Qualtrics, and employed quotas to obtain a sample that was nationally representative in terms of age, race/ethnicity, and geographic region. Using a sample recruited by the National Opinion Research Center (NORC), the remaining study features a nationally-representative probability sample from NORC's "AmeriSpeak Omnibus" survey.

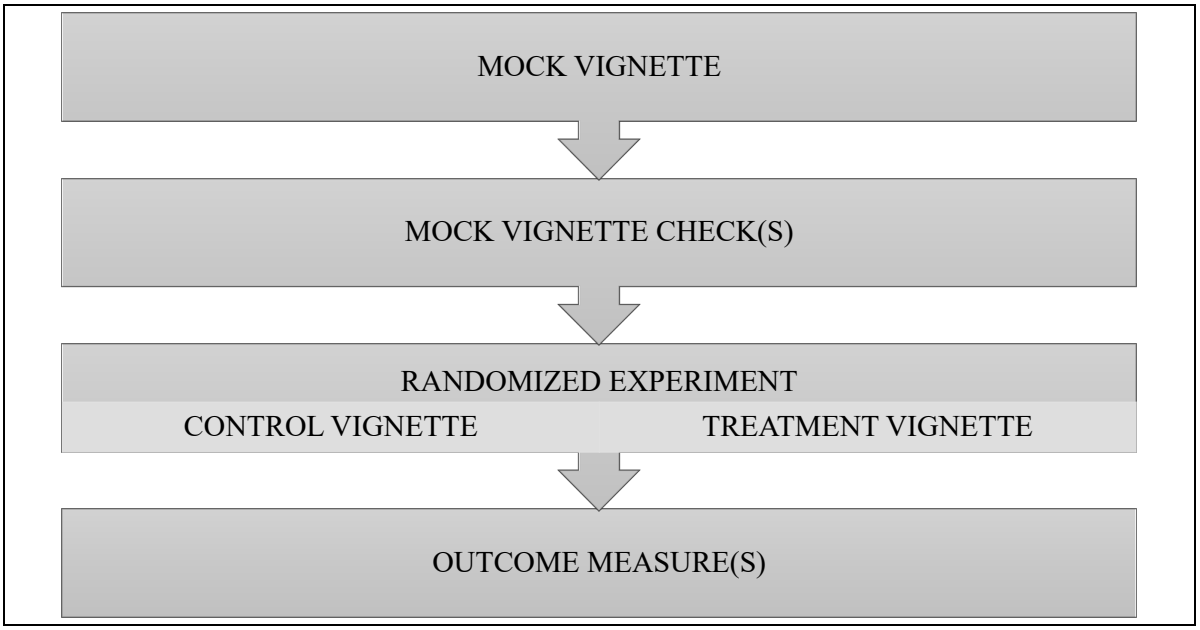
TABLE 1. Overview of Samples, Mock Vignettes, and Experiments

	MTurk 1 (n=603)	**Qualtrics** (n=1,040)	**NORC** (n=744)	**MTurk 2** (n=804)
Mock Vignette	Mandatory Sentencing	Mandatory Sentencing	Same-Day Registration	Scientific Publishing
Experiment Replicated	Student Loan Forgiveness	KKK Demonstration	Student Loan Forgiveness	Welfare Deservingness

Notes: Text for all mock vignettes and experimental vignettes appears in Supplemental Appendices B and C. "Student Loan Forgiveness" = Mullinix, Leeper, Druckman and Freese (2015); "KKK Demonstration" = Nelson, Clawson and Oxley (1997); "Welfare Deservingness" = Aarøe and Peterson (2014).

Each of these studies featured the same basic design depicted in Figure 1. Respondents in each study saw the same mock vignette (MV), then answered at least one factual question aimed at checking attentiveness to this MV. Again, we refer to this factual question as a mock vignette check (MVC).

FIGURE 1. Implementation of Mock Vignettes in Each Study



Notes: Design used in the MTurk 1, Qualtrics, MTurk 2, and NORC studies. Respondents in the Lucid study participated in this process twice. Each box represents a different screen viewed by respondents. In each study, timers were used on each screen to record the amount of time (in milliseconds) respondents spent on each screen, and a factual manipulation check (FMC) appeared after the outcome measure. All studies featured an experiment with two conditions.

Respondents were then randomly assigned to one of two conditions from a previously-published experiment (detailed below). After viewing one of these randomly assigned vignettes, respondents answered an outcome question(s) drawn from the original experiments that we replicated. Finally, in each study, we placed a factual manipulation check (FMC) immediately after the experimental outcome measure(s), and also featured timers on each screen. These latter steps permit benchmarking MVCs versus other attention measures (discussed in detail below).

In the fifth experiment—the Lucid sample study—each respondent experienced the same design featured in Figure 1 *twice*. In other words, within each of two separate rounds, every person taking the survey was randomly assigned one out of four possible MVs (and its corresponding MVCs), as well as one out of four possible experiments. (In the second round, respondents could not view the same MV, nor the same experiment, from the previous round.) This design has several advantages over

TABLE 2. Overview of Samples, Mock Vignettes, and Experiments (Lucid Study)

	Randomly Assigned Mock Vignette			
	1	**2**	**3**	**4**
Name of Mock Vignette	Scientific Publishing	Stadium Licenses	Sulfur Reductions	Plant Removal
	Randomly Assigned Experiment			
	1	**2**	**3**	**4**
Name of Replicated Experiment	Student Loan Forgiveness	KKK Demonstration	Welfare Deservingness	Immigration Policy

Notes: In the Lucid study, respondents were assigned to two rounds, each with one MV followed by one experiment (respondents could not be assigned the same MV or experiment twice). Text for all mock vignettes and experimental vignettes appears in Supplemental Appendices B and C. "Student Loan Forgiveness" = Mullinix, Leeper, Druckman and Freese (2015); "KKK Demonstration" = Nelson, Clawson and Oxley (1997); "Welfare Deservingness" = Aarøe and Peterson (2014); "Immigration Policy" = Valentino et al. (2019).

the previous studies. First, it permits estimation of correlations between different MVCs. Second, it effectively yields *two* observations per respondent, which improves model efficiency and statistical power. Third, because each experiment could be preceded by *any* of the four MVs, it enables us to investigate whether any observed results are dependent upon *which* particular MV was featured before the experiment. Table 2 lists the specific MVs and experiments featured in the Lucid study.

Mock Vignettes Featured

Regarding the MVs, the "Mandatory Sentencing" MV noted in Table 1 features adapted text from a published experiment by Gross (2008; see "Episodic Frame" on pp.185-86). All other MVs featured in Tables 1 and 2, however, were constructed by the authors, though were based upon actual research and/or recently published news articles (see Supplemental Appendix C for details). These latter MVs were one paragraph in length, and averaged approximately 140 words (min = 122;

max=159). In brief: "Same-Day Registration" discusses the potential costs of implementing same-day voting registration policies in various states; "Scientific Publishing" discusses controversy around a potential policy change in publishing of federally-funded research; "Stadium Licenses" discusses a small town's plan to produce an event license that would cover all stadium events; "Sulfur Reductions" discusses an upcoming requirement that ships reduce sulfur dioxide emissions; "Plant Removal" discusses a city council's new requirement that property owners remove hazardous vegetation from their properties. As an example, Table 3 provides the verbatim text of one of these MVs ("Scientific Publishing") and its corresponding MVCs. Verbatim text for all other MVs and MVCs can be found in Supplemental Appendix C.

Except for the "Minimum Sentencing" MV (which only had one MVC), each MV had three corresponding MVCs. Each MVC had between 5 to 6 closed-ended, randomized response options. By virtue of each MVC having only one correct answer, each MVC is coded as either incorrect (0) or correct (1). In every study except NORC, respondents were required to offer a response to each MVC, and in each study were not permitted to go back to a previously-viewed MV passage. The MVCs appeared in a fixed order, with later questions typically referencing material that appeared later in the MV's text.[6] When multiple MVCs were used, these were first coded as either incorrect (0) or correct (1), and then combined into an additive scale (see below).

[6] Particularly for the MVCs featured in the Lucid study, we made a concerted effort to keep the questions and response options similar in nature across each MV.

TABLE 3. Example Mock Vignette and Mock Vignette Checks (Scientific Publishing)

Mock Vignette	*A Passage from a Recent Magazine Article:* More than one hundred scientific societies and journal publishers are warning lawmakers not to move forward with a policy that would make all research supported by federal funding immediately free to the public. In three separate letters, they argue such a move would be costly, could bankrupt many scientific societies that rely on income from journal subscriptions, and would harm science in general. Lawmakers won't comment on whether they are actually considering a policy that would change publishing rules, and society officials say they have learned no details. But if the rumor is true, the order would represent a major change from current U.S. policy, which allows publishers to hold back federally-funded research from the general public for up to 1 year.	
Mock Vignette Check 1	*What was the topic of the magazine article you just read?*	(1) Literary Magazines (2) Scientific Research Publishing (3) Arts Funding (4) English Education (5) Immigration Policy (6) Funding for Space Exploration
Mock Vignette Check 2	*Regarding the rumored change in policy that was discussed, the magazine passage indicated that:*	(1) Lawmakers won't comment on whether they are considering the policy (2) Legal scholars stated the change in policy would be challenged in courts (3) Journal publishers have already begun preparing for the change in policy (4) Scientific researchers support the policy (5) All of the above (6) None of the above
Mock Vignette Check 3	*According to the magazine article you just read, current policy allows federally-funded research to be withheld from the general public for up to:*	(1) 1 Month (2) 6 Months (3) 1 Year (4) 3 Years (5) 5 Years (6) 10 Years

Notes: MVCs presented in this order. Response options (excluding "All of the above" and "None of the above") were randomized. Correct responses are highlighted in gray.

Prior Experiments Replicated

Regarding the experiments we featured (see Tables 1 and 2), the "Student Loan Forgiveness" study is a replication of an experiment conducted by Mullinix, Leeper, Druckman and Freese (2015). This experiment featured a control condition and a treatment condition, with the latter providing information critical of student loan forgiveness for college students. With support for student loan

forgiveness measured on a 7-point scale (ranging from *strongly oppose* to *strongly support*), the authors found that the treatment significantly reduced support for student loan forgiveness. This experiment has also been replicated successfully in previous research (e.g., Kane and Barabas 2019).

The "KKK Demonstration" study features the canonical experiment conducted by Nelson, Clawson and Oxley (1997). These authors found that framing an upcoming demonstration by the Ku Klux Klan as a matter of ensuring public order and safety, as opposed to a matter of free speech, yielded significantly lower public support for the demonstration to continue (again, measured on a 7-point scale ranging from *strongly oppose* to *strongly support*). Once again, this experiment has been replicated in prior studies (e.g., Mullinix et al. 2015; Berinsky, Margolis and Sances 2014).

The "Welfare Deservingness" study features the experiment conducted by Aarøe and Petersen (2014). To maintain only two conditions (as in the other experiments), we omitted the original control condition, leaving only the "Unlucky Recipient" and "Lazy Recipient" conditions. The authors found that, when discussing an individual as being out of a job due to a lack of motivation ("lazy"), as opposed to due to a work-related injury ("unlucky"), U.S. and Danish support for tightening welfare eligibility requirements ("for persons like him") significantly increases. This latter variable is referred to as "opposition to social welfare," and is measured on a 7-point scale (ranging from *strongly disagree* to *strongly agree*).

Lastly, the "Immigration Policy" experiment replicates an experiment, conducted in multiple countries, by Valentino et al. (2019). Again, to restrict the number of experimental conditions to two, we adapted the experiment to involve only two vignettes involving male immigrants: one is a "low-status" (i.e., low education and part-time working) Kuwaiti individual, and the other a "high-status" (i.e., highly educated and employed in a technical position) Mexican individual. The authors find that both lower-status individuals, and individuals from Muslim-majority countries, elicit lower public

support for allowing the individual to immigrate into the country. Specifically, the outcome measure is an additive scale comprising three separate items that gauge support for permitting the individual to work and attain citizenship in the respondents' home country. This scale ranges from 0 to 1, with higher values indicating greater support. Text for all vignettes, outcome response options, and factual manipulation checks can be found in Supplemental Appendix B.

RESULTS

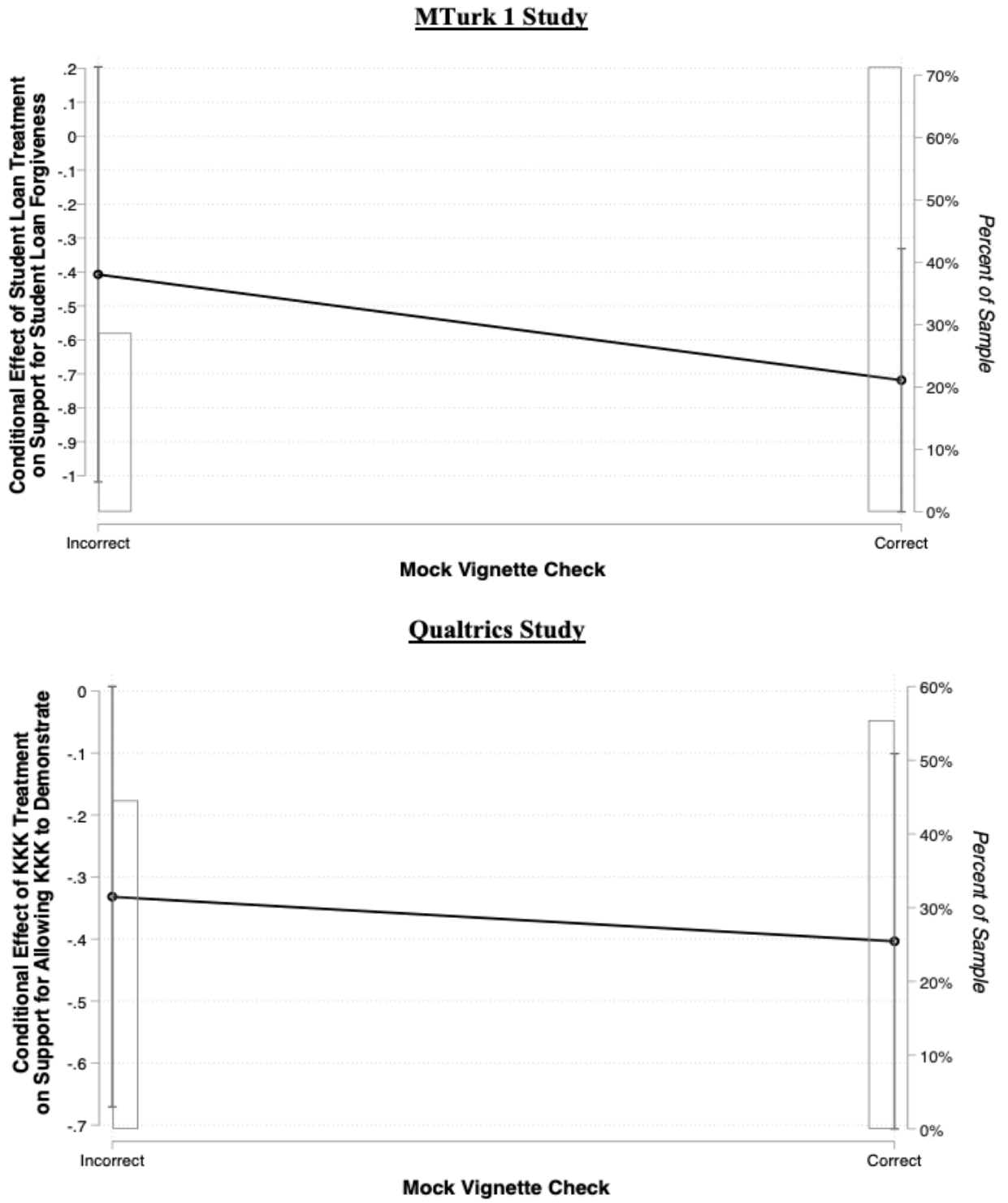
Beginning with performance on the MVCs, our MTurk 1 study obtained a passage rate (i.e., the share of respondents who answered the MVC correctly) of 71%, while our Qualtrics study obtained a passage rate of 55%.[7] For the NORC and MTurk 2 studies, which featured one MV with three MVCs, passage rates ranged for any given MVC ranged from 36% to 81%, and 44% to 80%, respectively. In the Lucid study, passage rates were generally between 50% and 80% (minimum=51%, maximum=82%). These rates are comparable to those from other techniques (e.g., IMCs and FMCs) as is the substantial variation in attentiveness across sample types (e.g., Kane and Barabas 2019).

MVC Performance and Treatment Effect Size

We now investigate whether MVC passage is associated with larger treatment effect sizes. The left *y*-axes of Figure 2 display the estimated treatment effects in the MTurk 1 study (top panel) and Qualtrics study (bottom panel), among (1) those who answered the MVC incorrectly (non-passers) and

[7] Observing a relatively higher level of attentiveness in the MTurk samples is consistent with research by Hauser and Schwarz (2016). MTurk respondents also adhere to other protocols, such as not "cheating" as much on political knowledge questions relative to subjects from other platforms (Clifford and Jerit 2016).

FIGURE 2. Mock Vignette Check Passage Associated with Larger Treatment Effects



Notes: Figure displays treatment effects for "Student Loan Forgiveness" experiment (top panel) and "KKK Demonstration" experiment across performance on the mock vignette check question (95% CIs shown). In both studies, only one MVC was featured. Total N = 603 (MTurk Study 1) and 1040 (Qualtrics). Histograms display the percent answering incorrectly or correctly (see right *y*-axis).

(2) those who answered the MVC correctly (passers). Histograms are also featured to indicate the share of passers and non-passers in each study (see right *y*-axes).

Beginning with the MTurk 1 study, the estimated treatment effect does indeed increase in magnitude as we move from MVC non-passers (29% of sample) to passers (71% of sample). Among MVC non-passers, the treatment effect is a decrease of .41 for support for student loan forgiveness (from 4.94 in the control condition to 4.53 in the treatment condition), and was non-significant ($p=.15$). Among MVC passers, however, the estimated treatment effect is a decrease of .72 (from 5.13 to 4.41), which was significant at the $p<.001$ level. This difference in treatment effects represents a 76% increase in effect size and, as revealed by a difference-in-differences (DID) estimate (not shown), is statistically significant at $p<.05$ (one-tailed). Lastly, the treatment effect for the sample as a whole (i.e., the ITT) is equal to -.63, which is substantively smaller than the estimate among those who passed the MVC (-.72).

For the Qualtrics study, we again observe a stronger treatment effect among MVC passers (64% of sample) versus non-passers (36% of sample). Among passers, the treatment effect of the "Public Order" (versus "Free Speech" frame) is a decrease of .51 in support for allowing the KKK to demonstrate (from 3.15 to 2.65), which was significant at the $p<.01$ level. However, among non-passers this decrease is only .36 (from 2.83 to 2.47), and was not significant at the $p<.05$ level. Thus, going from MVC non-passers to passers yields a 41% increase in effect size, though, in this case the DID was not quite statistically significant ($p=.15$). With the treatment effect for the sample as a whole being equal to -.46, this study, like the previous one, illustrates how neglecting to account for inattentiveness will tend to yield weaker treatment effect estimates.

Overall, both the MTurk 1 and Qualtrics studies provide preliminary evidence that better performance on MVCs is associated with stronger treatment effects. Further, these analyses exemplify

how researchers can use the mock vignette performance in their own analyses: results are displayed for the sample as a whole, but, as an additional test of the hypothesis that accounts for respondent inattentiveness, results are also displayed for only those respondents who passed the MVC.[8] If sample inattentiveness is systematically attenuating the treatment effect, then the researcher should observe a treatment effect estimate larger in magnitude when analyzing only those who passed the MVC.

Compared to the previous studies, a major advantage of the NORC and MTurk 2 studies is that, while each features only one MV, there are *three* accompanying MVCs. Having multiple MVCs is likely to yield a scaled measure of attentiveness that contains less measurement error than that of a single MVC (e.g., see Ansolabehere, Rodden, and Snyder 2008).

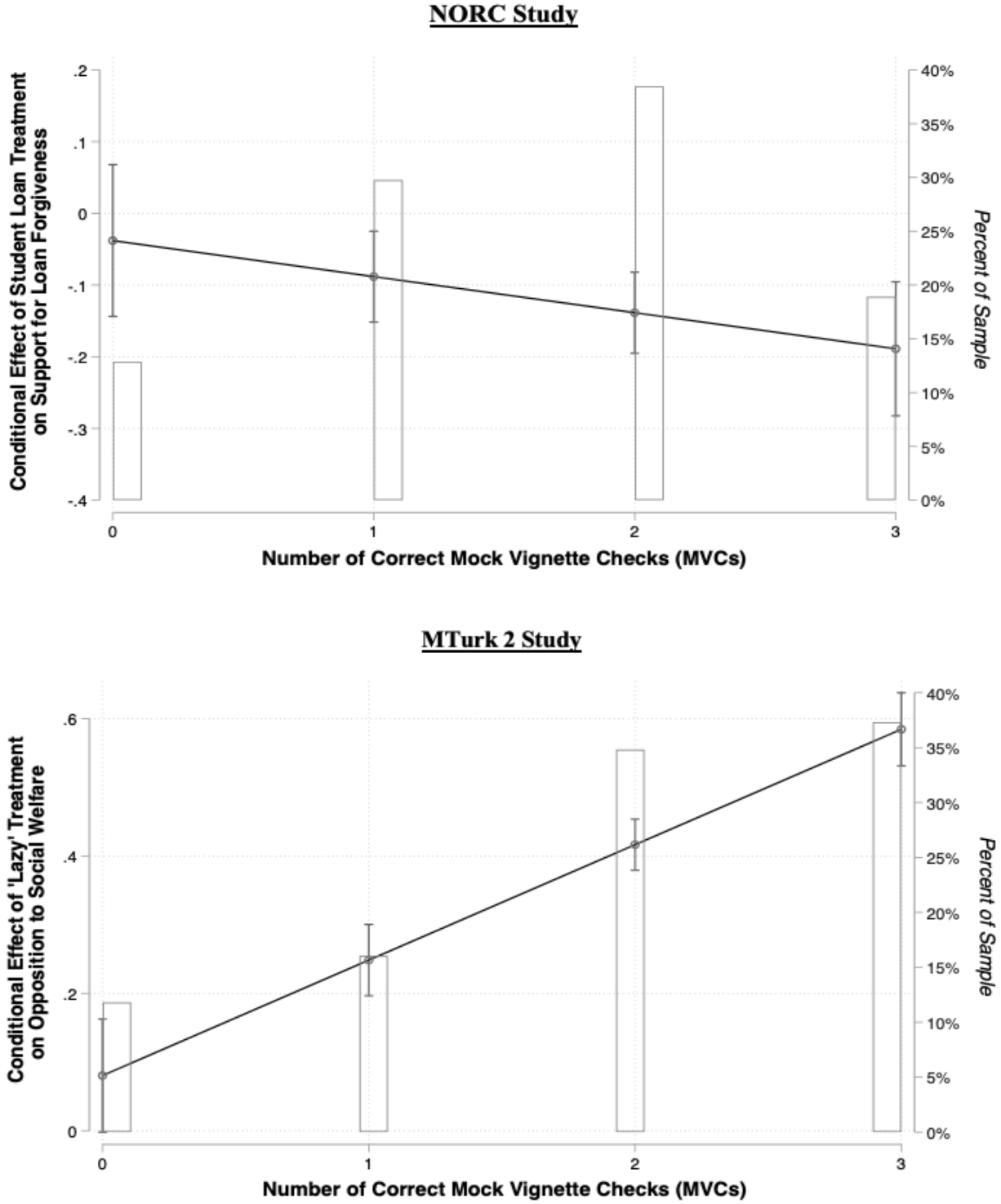
Figure 3 displays the conditional average treatment effect (CATE), in both the NORC (top panel) and MTurk 2 (bottom panel) studies, across performance on the MVCs. In each study, the dependent variable has been rescaled to range from 0 to 1 to enhance interpretability. Underlying histograms are again featured to convey MVC performance distribution in each study, with the right *y*-axis displaying the percentage of the sample passing a given number of MVCs.

Turning first to the NORC study (top panel of Figure 3), which featured the "Student Loan" experiment (Mullinix et al. 2015), we indeed see that whereas the estimated CATE is only slightly negative (-.038, or -3.8 percentage points) and non-significant among those who passed 0 MVCs, the estimated CATE grows substantially more negative, and becomes statistically significant (i.e., the 95% CIs no longer overlap with 0), with better performance on the MVCs, thus replicating the negative treatment effect observed in the original study.[9] This interaction between treatment and MVC

[8] This latter estimand is therefore akin to the average effect of receipt for compliers (AERC [see Harden, Sokhey, and Runge 2019 Supplmental Appendix pp.10-11]).

[9] For the NORC sample as a whole, the estimated intent-to-treat (ITT) effect = -.12 ($p<.001$).

FIGURE 3. Mock Vignette Check Performance Associated with Larger Treatment Effects



Notes: Figure displays treatment effect estimates for "Student Loan Forgiveness" experiment (top panel) and "Welfare Deservingness" experiment across performance on the mock vignette check scale (95% CIs shown). Total N=744 (NORC) and 804 (MTurk Study 2).

performance was statistically significant ($p<.05$, one-tailed). At the highest level of MVC performance (all three MVCs correct, approximately 19% of the sample), the estimated CATE reveals a nearly 20 percentage-point decrease in support for student loan forgiveness. This effect is far larger than the -3.8 effect that was observed among those who did not answer any MVCs correctly (approximately 13% of the sample). As this analysis further demonstrates, inattentiveness in the sample attenuates the treatment effect observed for the sample as a whole, thereby increasing the risk of a Type II error and undermining hypothesis testing.[10]

The results for the MTurk 2 study (see bottom panel of Figure 3) are even more pronounced. Replicating the social welfare deservingness experiment (Aarøe and Petersen 2014), the bottom panel of Figure 3 indicates that the effect of the "lazy" treatment on opposition to social welfare substantially increases with better MVC performance. This interaction between treatment and MVC performance was again statistically significant ($p<.001$). Specifically, at 0 MVCs correct (approximately 12% of the sample), the estimated treatment effect is relatively small (.08 on a 0-1 scale), with a 95% confidence interval that narrowly overlaps with 0. However, at 3 correct MVCs (approximately 37% of the sample), this estimated treatment effect increases in size by *more than sevenfold* to .58. Again,

[10] As a further illustration of this point: we observe that, among the 140 respondents who passed all 3 MVCs, the effect was .23 ($p < .01$) and power equaled .96 (two-tailed, alpha=.05). Conversely, given the weak effect size among those who passed 0 MVCs (effect=.076, se=.078), to have power equal .96 among these respondents we would have required a sample size of 1,414, or over 700 in each arm of the experiment.

inattentiveness among some respondents yielded substantially weaker treatment effects for the sample as a whole.[11]

Because this latter set of analyses involved an MVC scale rather than a single binary measure, these results exemplify a second way in which researchers can use mock vignettes in their analyses: after reporting the estimated treatment effect for the sample as a whole, researchers can specify an interaction between the treatment variable and the MVC performance scale. In essence, this enables the researcher to investigate the degree to which the estimated treatment effect increases in magnitude across MVC performance, and without jeopardizing the study's internal validity via introducing post-treatment bias. Finding that the estimated treatment effect increases in magnitude at higher levels of MVC performance, for example, would indicate that inattentiveness in the sample partially undermined one's hypothesis test, and thus serve as more robust evidence in favor of a hypothesis. This approach can also be especially helpful as a diagnostic tool for researchers who obtain null results for a given experiment: if no such change in treatment effect magnitude is observed across MVC performance, this would suggest an ineffective manipulation, or an incorrect underlying theory, rather than a problem arising from sample inattentiveness.

We now turn to the Lucid study, our last set of experiments, in which each respondent participated in two rounds. In each round, respondents were randomly assigned to one of four MVs and randomly assigned to one of four experiments (each with a randomly assigned control and treatment condition). First, we present results from a "grand model" that estimates CATEs using data from the full set of experiments and MVs to gauge the average performance of the mock vignette technique. We next subset the data by MV, and show how CATEs vary as a function of MVC

[11] For the MTurk 2 sample as a whole, the estimated intent-to-treat (ITT) effect = .41 ($p<.001$).

performance. Using additional models, we then probe whether our MVs are relatively interchangeable or, conversely, particular MVs outperform others.

Table 4 displays the results from a linear model with standard errors clustered by respondent.[12] The model takes the following form:

$$Y_{ir} = \alpha_{ir} + \beta_1 T_{ir} + \beta_2 MVC_{ir} + \beta_3 T_{ir} \times MVC_{ir} + \epsilon_{ir}$$

where *i* indexes individuals, *r* indexes rounds, *Y* represents the outcome measured in terms of control group standard deviations within each experiment, *T* is a treatment indicator and MVC represents the number of correct MVCs. We assess the robustness of the linearity assumption in Supplemental Appendix E, and find that the data are consistent with a linear multiplicative model.

As shown in Table 4, the interaction between treatment status and MVC performance is statistically significant ($p < .001$). At 0 correct MVCs (approximately 22% of the sample), the CATE is 28% of a standard deviation. This corresponds to approximately a .50 scale point shift on a 7-point Likert scale.[13] However, at 3 correct MVCs (41% of the sample), the CATE is approximately 2.7 times larger, reflecting a 76% standard-deviation (or 1.50 scale point) shift in the outcome variable.

[12] Fixed effects and random intercept models were also estimated. However, this does not produce any substantive differences in estimates because the grouping variables are uncorrelated with treatment status due to random assignment.

[13] Given the need to aggregate across multiple studies with different outcome measures, we standardize our outcomes using control group standard deviations. However, three out of the four experiments feature seven-point Likert scales with standard deviations approximately equal to 2, and thus, we also report raw scale quantities to facilitate the interpretation of effects. Though the immigration study did

TABLE 4. Conditional Effect of Treatment on Outcome across MVC Passage Rates

	Experimental Outcome Measure
Treatment Status	.279***
	(.036)
Mock Vignette Check Score	-.033***
	(.012)
Treatment Status × Mock Vignette Check Score	.162***
	(.017)
N	11,056

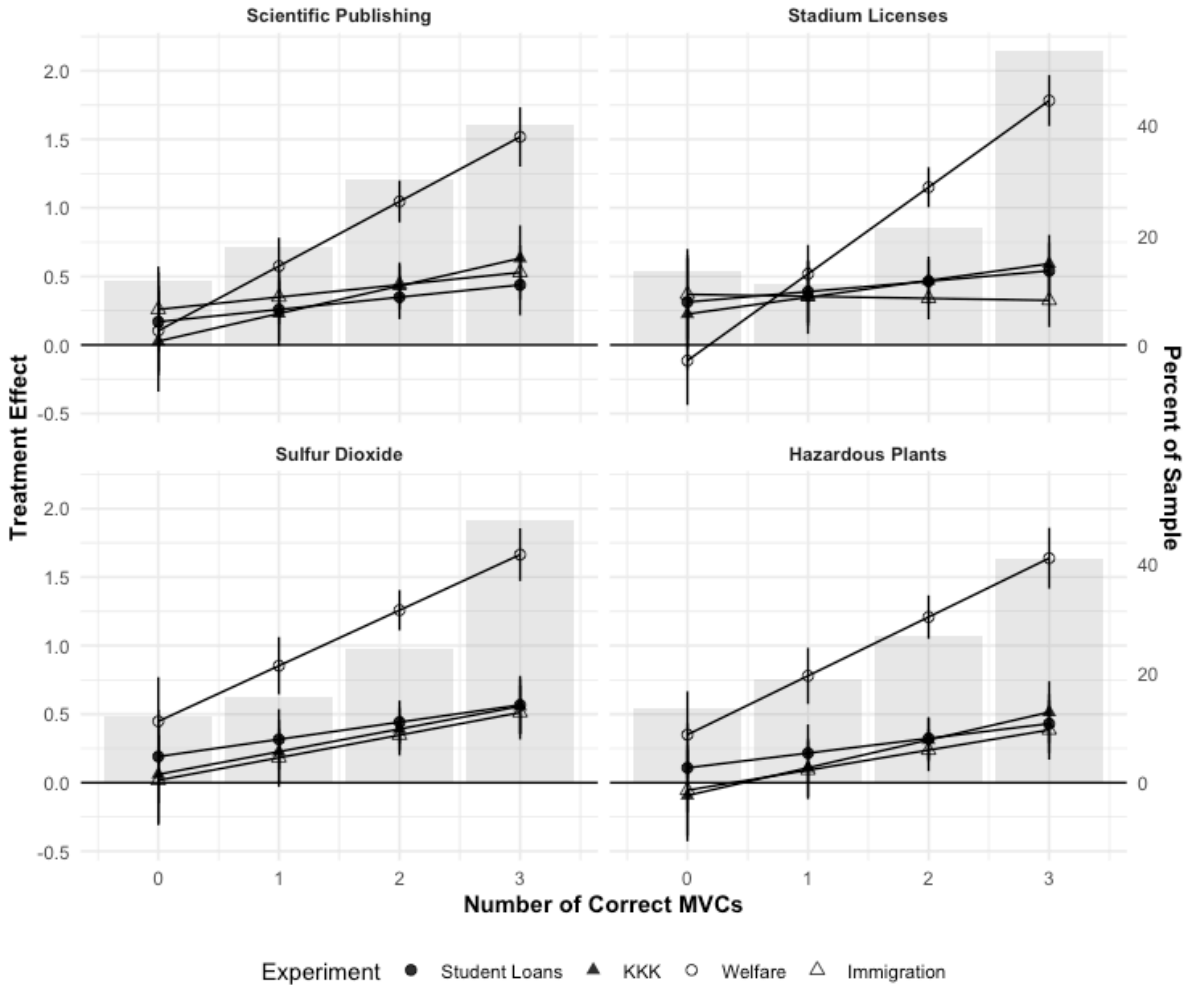
Notes: Lucid study. OLS regression coefficients with standard errors clustered by respondent. Outcome is standardized within each experiment (control group standard deviations). Mock Vignette Check Score ranges from 0 to 3. ***p<0.001 (one-tailed hypothesis tests).

To provide a visual sense of how CATEs vary as a function of MVC performance within each MV-experiment pair, we present CATE estimates for each MV and experiment in Figure 4. This figure reveals that MVC performance is positively associated with CATEs in 15 out of the 16 MV-Experiment combinations.[14] The relationship between MVC performance and CATEs is strongest in the *Welfare* experiment. This is likely a function of the large ITT estimate for that experiment, which is a 1.17 standard-deviation shift in the outcome variable (approximately 2 scale points), whereas ITT estimates for the other experiments range from .34 to .41 standard deviations (70 - 80% of a scale point on a 7-point Likert scale). Moreover, CATEs among those who perform worst on the MVC are not

not use a single seven-point Likert scale, outcomes were measured using three items that sum to a score of seven. The standard deviation for this study is 1.82.

[14] CATE estimates among those assigned to the stadium licenses-immigration pair decrease as a function of correct MVCs. However, differences between those who score the lowest and highest in terms of MVC performance are not statistically distinguishable.

FIGURE 4: CATE Estimates Across Experiments (by Mock Vignette Featured)



Notes: Conditional average treatment effects across number of correct MVCs for each mock vignette-experiment pair. Points represent CATE estimates (95% CIs shown). Histogram represents the percent of the sample correctly answering *x* MVCs.

statistically discernible from zero in all but three out of sixteen cases, whereas they are statistically significant in *every* case among those who answered all MVCs correctly.

Figure 4 also suggests that the relationship between MVC performance and CATEs is relatively similar regardless of the particular MV that is used. We conducted an explicit test of this possibility (see Supplemental Appendix F for details), and find differences between MVs—in terms of predicting larger CATEs—to be minimal and not statistically discernible from zero.

MVC Performance and Other Measures of Attentiveness

The previous section finds strong and consistent evidence that larger treatment effects are observed among those who perform better (versus worse) on our proposed measure of attentiveness (i.e., MVCs). An implication of these findings is that MVC performance should be associated with better performance on *other* measures of attentiveness. In other words, attention to the MV (as measured by performance on MVCs) should be associated with measures of attentiveness to the experiment itself. We first note, however, that performance on a given MVC generally had substantial and statistically significant pairwise correlations with performance on *other* MVCs. For example, the Lucid study MVCs had pairwise correlations ranging from .55 to .63 ($p<.001$), and Cronbach's alpha (α) values ranging from .60 to .74.[15] Further, in the Lucid study, the pairwise correlation between round 1 and round 2 MVC performance was quite strong at .60 ($p<.001$) and did not vary much across MVs (.55 to .63). Such correlations are considerably higher than those found for instructional manipulation checks (IMCs, also known as "screeners"; (Berinsky, Margolis, and Sances 2014; Hauser, Paolacci, and Chandler 2018; Thomas and Clifford 2017), as well as correlations found between alternative measures of attentiveness (Niessen, Meijer, and Tendeiro 2016).

Question timers represent an alternative means by which to measure attentiveness to an experiment. The logic here is that, in general, less time spent on an item is indicative of less attentiveness to its contents (Niessen, Meijer, and Tendeiro 2016; Wood et al. 2017). We implemented

[15] The "Same-Day Registration" MVCs displayed noticeably smaller, though still positive and statistically significant, pairwise correlations (ranging from .11 to .32, $p<.01$) and $\alpha=.40$. This may be partly due to NORC respondents being permitted to skip MVCs (which was recorded and counted as "incorrect" (0)).

question timers on each mock vignette, as well as on every screen of the experiment: the randomly assigned vignette, the outcome measure, and the FMC. Consistent with previous research (e.g., Wood et al. 2017), we log-transform each timer item, and subsequently regress it onto MVC performance, yielding an estimate of the percent change in time spent on a given item per a one-unit increase in MVC performance.

We present the full results of our analyses in Supplemental Appendix D. To summarize results for the MTurk, Qualtrics and NORC studies, better performance on the MVC consistently predicts greater latency (i.e., time spent) on (1) the mock vignette itself, (2) the experimental vignettes, and (3) the experiment's outcome measure. These differences were positive in sign and statistically significant at $p<.05$ or below in all but once instance.[16] For example, in the KKK experiment, passing (versus failing) the MVC predicts 132% more time spent reading the "free speech" vignette. (In terms of raw times, MVC non-passers spent an average of 28 seconds while passers spent an average of 68 seconds.) Further, in every case, those who passed the MVC spent significantly more time on the survey itself.[17] In every single Lucid experiment, better MVC performance predicts significantly greater time spent

[16] The one instance is that of time spent on the outcome measure in the NORC study, for which the estimated difference was small and non-significant.

[17] Relatedly, as a means of ensuring data quality, Qualtrics independently flags respondents with unusually fast survey completion times (i.e., "speeders"). In the Qualtrics study, 36 percent of MVC non-passers were flagged as a "speeder", whereas only 11 percent of passers were flagged as such. Given that Qualtrics would normally exclude these "speeders" from one's sample, all other analyses with Qualtrics data exclude these "speeders".

on a given timer. Thus, in 39 out of 40 separate tests, we find that better MVC performance is associated with significantly more time spent on experimental items.

Lastly, given that factual manipulation checks (FMCs) are designed to measure individual attentiveness to the actual experiment's vignettes, we also find a remarkably strong relationship between MVC performance and passing the FMC: MVC performance predicts anywhere between a 35 (Qualtrics and NORC) and 49 (MTurk 1) percentage-point increase in likelihood of correctly answering the experimental FMC. In the Lucid study, these effects were even stronger, ranging from 41 to 68 percentage points. Thus, in 8 out of 8 separate tests, we find that better MVC performance is associated with significantly greater likelihood of correctly answering a question about the contents of the experiment.

Does Using MVs Significantly Alter Treatment Effects?

The previous sections offer consistent support for using MVCs as a means of measuring respondent attentiveness and for examining treatment effects among those likely to have been attentive to one's experiment. However, a natural question is whether the act of featuring an MV, in and of itself, yields an ITT estimate for the experiment that is substantially different from what would have been observed had no MV been featured. For example, the MV might prime various considerations that would not have otherwise been primed, potentially rendering respondents more, or perhaps less, receptive to the treatment (on average). Alternatively, as the MV supplies an additional quantity of information, and MVCs constitute additional demands upon respondents' cognitive stamina, perhaps featuring an MV results in greater respondent fatigue and, thus, weaker treatment effects.

To investigate this potential concern, we designed the Qualtrics and Lucid studies such that a random subset of respondents was selected to not receive any MV prior to the experiment. This enables

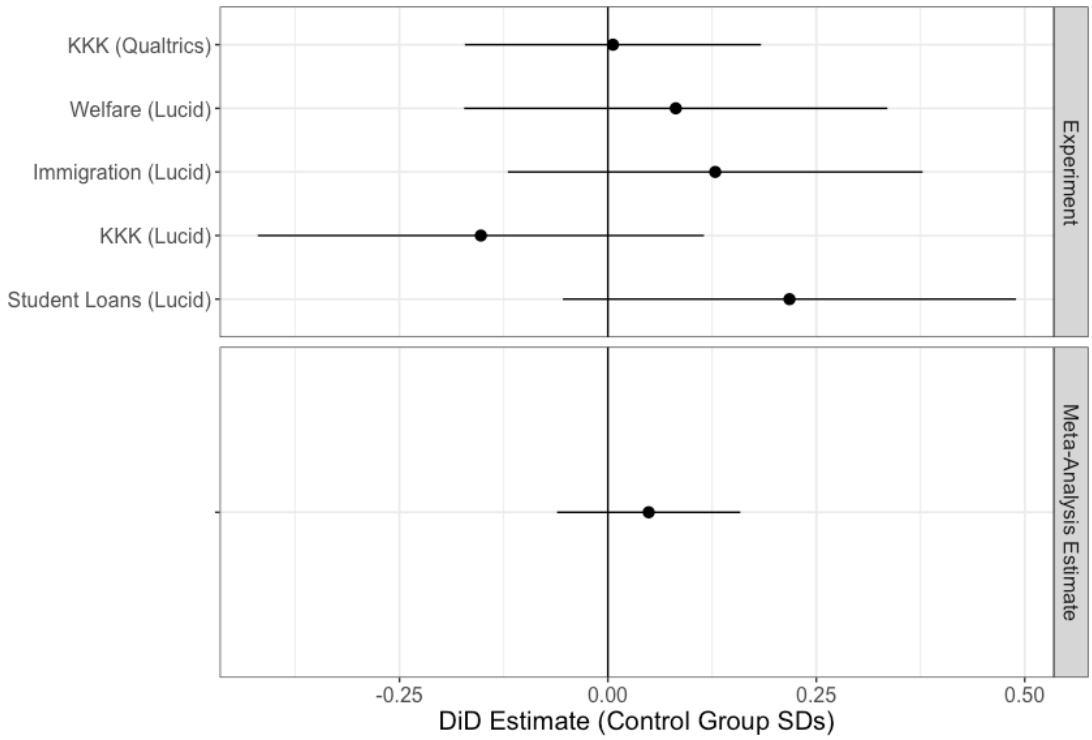
us to directly investigate whether the experimental treatment effects are substantially different for those who did, versus did not, view an MV (and answer MVCs) prior to the experiment.

The results of this investigation appear in Figure 5. Beginning with the Qualtrics study, wherein 25% of the sample was not shown an MV (n=256), there is no statistically distinguishable difference in treatment effect estimates between those who observed and did not observe the MV.

In the Lucid study, 20% of the respondents in the first round (n=1000) were randomly selected to not receive an MV. We therefore examined whether, within the first-round experiments, exposure to an MV yielded significantly different treatment effects in any of the four experiments. This effectively amounts to four additional tests of whether featuring an MV alters treatment effects. As per the figure, we find no evidence that those respondents who observed, versus did not observe, an MV before the experiment exhibited significantly different treatment effects. Treatment effects were, in each experiment, substantively and statistically similar across these two groups. In particular, the DID estimates are statistically indistinguishable from zero in all four cases.[18] Moreover, the sign on the DID estimate was not consistent—that is, in one instance the sign is opposite the ITT estimate (the KKK study), but in the other three instances the sign is the same as the ITT estimate. Thus, in addition to there being no significant interaction, there is also no consistent pattern with respect to whether featuring an MV attenuates or augments treatment effects. Finally, when we compute a meta-analytical summary estimate of the effect size across all of these studies using random-effects meta-analysis, we

[18] Each of these four analyses had between 1239 and 1270 respondents in total, making it unlikely that such results are simply due to insufficient power.

FIGURE 5: No Significant Change in Treatment Effects When a Mock Vignette Is Used



Notes: Figure shows the difference-in-differences (DiD) estimate for experiments with and without a preceding mock vignette. Points represent DiD estimates (95% CIs shown). Top panel presents individual estimates, whereas bottom panel presents the random-effects meta-analysis estimate computed by the R package *rmeta*.

find that the "MV inclusion effect" is small (.05 control group standard deviations) and also not statistically distinguishable from zero.[19]

[19] In contrast to fixed-effect meta-analysis, which assumes that studies are estimating a single "true" effect, random-effects meta-analysis models assume that effects are drawn from a larger population, and may vary from study to study. In our case, the fixed-effect meta-analysis estimate (.047) is very similar to the estimate obtained by assuming random effects.

Mock Vignette Placement

Another advantage of the Lucid design is that, for any given respondent, there is variation in the *placement* of the MV relative to the experiment. While we contend that, to avoid post-treatment bias, MVs should appear *prior to* the researcher's experiment, it is an open question as to whether researchers would benefit most from placing the MV directly before (versus long before) their experiments. We therefore investigated whether the *treatment* X *MVC* interaction (i.e., the CATE) increases in magnitude as a result of the MV appearing directly (versus long) before the (second-round) experiment (see Supplemental Appendix F for details). We find that the difference in DID estimates when the MV appears directly before (versus long before) the treatment is 0.7% of a standard deviation ($p = .91$). Thus, while we find that CATEs were slightly larger when MVs were placed directly before the treatment, the effect is small and not statistically discernible from zero. This suggests that MVs do not necessarily need to appear immediately before one's experiment to adequately capture attentiveness. However, we caution that this result may be partly because the two MVs were placed relatively close together (i.e., (in)attentiveness was likely similar, for any given respondent, at both points in time in our study). As such, while an MV placed long before the survey may also suffice, we nevertheless recommend placing MVs directly before experiments given (1) the lack of evidence for a priming/fatigue effect (noted above), and (2) the underlying goal of measuring attentiveness to the experimental portion of the survey.[20]

[20] In this vein, we do find that the correlation between timers and MVC performance in round 2 correlate (slightly) more strongly with timers and FMCs on the round 2 experiment than did timers and MVC performance in round 1, suggesting that attentiveness levels shortly before (versus longer before) the experiment more closely resemble attentiveness during the experiment.

Demographic Patterns in MVC Performance

A common issue with analyzing attentive respondents is that the subset of attentive respondents in one's sample may differ from the broader sample on a variety of demographic and politically-relevant variables (e.g., Thomas and Clifford 2017). Such a difference is likely insofar as attentiveness is not randomly distributed in the population. To this end, for each of our studies, we ran a single regression model wherein we regressed MVC performance onto the following variables (all measured pre-treatment): *gender, race, age, income, education, political interest, party identification, and ideological self-placement*.

Due to spatial constraints, the full set of results can be found in Supplemental Appendix G. Overall, the only variables showing a consistently sizable and significant ($p<.10$) relationship with MVC performance across the five studies were (1) race, and (2) age. Specifically, non-White respondents tended to have lower MVC performance relative to Whites (generally on the order of 10 to 20 percentage points) and older respondents tended to perform substantially better than younger respondents (e.g., in the Qualtrics study, which displayed the strongest relationship between age and MVC performance, moving from the 25th to 75th percentile in age predicted an 18 percentage-point improvement in MVC performance), though no significant effect was found for age in the NORC study. These patterns are consistent with those of other studies wherein researchers analyzed an attentive subset of their sample (see Thomas & Clifford (2017, 192)). Overall, however, correlations between these demographic variables and MVC performance were substantively modest in size. Age, for example, correlates with performance at .33 or less across all of our studies. In the NORC study, which saw the largest effects for race (i.e., African-American and Hispanic identification) on MVC performance, the pairwise correlations were $<=|.15|$. Further, when analyzing MVC passers versus the sample as a whole (in the studies with only one MVC (MTurk 1 and Qualtrics)), the sample

composition does not substantially change. The average age among MVC passers in the Qualtrics study, for example, is 49, while it is 46 for the sample as a whole.

Importantly, we do not find any consistent effects for education, *nor do we find consistent effects for any political variables*—i.e., political interest, party identification, and ideological self-placement. This latter finding, in particular, helps assuage the potential concern that only, for example, highly educated and/or politically interested respondents will be able to successfully answer MVCs.

Thus, as prior studies have duly noted (Berinsky, Margolis, and Sances 2014; Thomas and Clifford 2017), we caution that analyzing the attentive may alter the demographic composition of the sample. Researchers can be transparent about this concern by, for example, noting correlations between demographic predictors and MVC passage, and/or (if subsetting on MVC performance) noting changes in the demographic composition of the sub-sample relative to the initial (full sample) analysis.[21]

That said, it is worth emphasizing that, with a pre-treatment measure of attentiveness, any relationship between demographic variables and attentiveness is a concern not about statistical bias but, rather, sample representativeness and the generalizability of one's findings. In other words, even if the attentive sub-sample differs demographically from the full sample, this potentially poses

[21] Another option would be to control for an interaction between treatment and the demographic variable that is highly predictive of MVC performance (assuming the demographic variable is measured pre-treatment). We performed this procedure for our MTurk 2, NORC and Lucid studies (which featured a continuous MVC scale), and found only minor changes in CATE size, and no substantive change in *p*-values for the CATE whatsoever. Again, researchers should be fully transparent about this modeling choice, noting differences in results with and without this control specified in the model.

constraints on the external validity of the results, rather than threatening the survey experiment's internal validity. Notably, however, existing research finds remarkably homogeneous treatment effects across samples with substantially different demographic compositions (e.g., Coppock, Leeper, and Mullinix 2018; Mullinix et al. 2015). Conversely, neglecting to account for inattentiveness whatsoever risks obtaining treatment effect estimates that are downwardly biased, potentially yielding null findings.

Detecting Significant Effects Among Attentive Sub-Samples

As implied in the previous section, analyzing respondents with relatively better MVC performance means analyzing a *subset* of one's sample. This raises practical questions concerning statistical power and, specifically, whether one can still detect statistically significant treatment effects when analyzing this attentive sub-group. We investigate these concerns in each of our studies (see Supplemental Appendix I for details). To summarize the results, because we consistently a find a larger treatment effect among the more attentive, we find that this helps to offset the loss of power that arises from subsetting the sample on MVC performance. In fact, in some cases we obtain a *larger t*-statistic on the treatment effect among the attentive sub-sample. Yet even in the cases where the treatment effect *t*-statistics decrease in magnitude, our results consistently show that the researcher can nevertheless uncover a statistically significant treatment effect (i.e., $p<.05$) even among the most attentive sub-sample of respondents.

Finally, to serve as potential guidance for researchers, we summarize in Table 5 how we constructed and implemented MVs and MVCs across our studies. This table also documents how MVC performance can be incorporated into one's analysis, as well as how results can be presented in a transparent fashion.

TABLE 5. Summary of Mock Vignette (MV) and Mock Vignette Check (MVC) Protocols

Construction	Mock Vignettes (MVs) were relatively short (approx.. 140 words), and did not contain obvious partisan content (e.g., references to well-known political figures, parties, or highly contentious policies). Mock Vignette Checks (MVCs) were designed to be relatively simple to answer if one paid attention to the vignette. For example, the correct response options use language that is verbatim to the language in the corresponding MV. In most of our MVs that used multiple MVCs, the first MVC asked about the broad topic. Subsequent MVCs asked about specific content featured earlier or later in the MV.
Implementation	The MV and MVC(s) were placed immediately before our experiment of interest. MVC(s) immediately followed the MV, appearing on a separate screen. Each MVC appeared on a separate screen with no ability to go backward or (in all but one study) skip over the question. MVCs had at least 5 (randomized) response options to minimize respondents' ability to correctly guess the MVC answer. Factual manipulation checks (FMCs) and timers on the experimental vignettes were used to confirm that MVC performance correlates with attention to the experiment.
Analysis	In the interest of full transparency (and as done in our study), researchers should report treatment effect (ITT) among full sample before incorporating MVC performance. To increase transparency, researchers can also report passage rates for MVC item(s), as well as any substantive demographic changes to the sample when analyzing those who answered correctly (versus the sample as a whole). Respondents were subsetted into varying levels of attentiveness based upon MVC performance; interactions between treatment and MVC performance permitted statistical analysis of treatment effect sizes at higher (versus lower) levels of attentiveness. Stronger treatment effects among those who were more (versus less) attentive are taken to constitute relatively stronger evidence against the null hypothesis.

Note: Summary of how MVs and MVCs were constructed and implemented across our studies, and recommendation for incorporating MVs and MVCs into one's analysis.

DISCUSSION & CONCLUSION

The growth of experimental social science has exploded in recent years due to technological advances that allow survey experiments to be fielded online. However, a persistent challenge arising from this mode of research is that of respondent inattentiveness, which stands to bias treatment effects downward. In this paper, we proposed mock vignettes (MVs) as a technique that enables scholars to assess treatment effects across varying levels of attentiveness without inducing post-treatment bias. First, we showed that conditional average treatment effects increase in magnitude with better performance on mock vignette checks (MVCs) across five studies. We then provided evidence that MVC scores correlate substantially with other measures of attentiveness, such as factual manipulation checks and response latencies (i.e., timers). Next, we found that featuring an MV does not have any discernible effect on experimental results and that our MVs were essentially interchangeable in predicting CATEs. Further, we found that while some demographic characteristics (particularly age and race) are associated with MVC performance, these correlations were substantively small, and that political variables such as partisanship, ideology, and political interest do not significantly predict MVC performance. Finally, we found that analyzing a relatively attentive subsample does not necessarily undermine a researcher's ability to uncover significant treatment effects at conventional levels of statistical significance.

In our Supplemental Appendix C, we provide text and performance analytics for a variety of pre-tested MVs that can be used by researchers. If scholars wish to use them or construct their own, we highlight the following suggestions based upon our studies' designs (see also Table 5). First, MVs ought to present subjects with a vignette that is broadly similar in nature to the kind of content featured in the experiment itself, but that is unlikely to have an effect on the outcome. The latter point is important, given the possibility of spillover effects in survey experiments (Transue, Lee, and Aldrich

2009). Second, as per Table 5, we recommend that scholars present MVCs as forced response questions to avoid missing data, and with no back button to prevent the possibility of looking up answers to the MVC. Third, as with all measures of attentiveness, we expect that MVCs will inevitably contain some degree of measurement error. Thus, multiple-item scales are advisable where possible. Finally, following Berinsky, Margolis and Sances (2014), we urge researchers to be fully transparent by presenting the ITT for the sample as a whole before re-estimating the treatment effect on those deemed to be attentive, or analyzing whether (and to what degree) the treatment effect increases in magnitude as attentiveness increases.

Though we identify several distinct advantages to the MV approach, its use does not obviate the need for other tools that gauge attentiveness, such as manipulation checks. Manipulation checks, particularly treatment-relevant factual manipulation checks (FMC-TRs) and subjective manipulation checks, provide information about the degree to which experimental manipulations were perceived and efficacious, respectively. These measures remain essential for describing sample characteristics, determining whether the experimental manipulation is affecting the theorized independent variable of interest, and diagnosing whether and to what extent inattentiveness is affecting one's results. Moreover, if the necessary assumptions for identifying complier average causal effects (CACEs) hold, such measures can be used in an instrumental variable (IV; e.g., 2SLS) setting. As we note above, while our approach and the IV approach differ substantially in terms of implementation and, ultimately, recover different estimands, both can potentially be used in parallel.[22]

[22] We re-analyzed the data of our Lucid study experiments using the IV (specifically, 2SLS) approach as recommended and implemented by Harden, Sokhey, and Runge (2019). Notably, we find that CACE estimates vary dramatically depending upon the latency cut-off that is employed, sometimes

Moving forward, we note that, as MVs are text-based vignettes, it remains unclear to what extent the MV approach will be effective for survey experiments that involve non-textual visual and/or auditory stimuli (e.g., photos, videos, or sound recordings). We believe this presents a useful avenue to explore in future research. More broadly, we encourage scholars to further investigate the nature of inattentiveness in experiments. We speculate that, in the aggregate, our MVC measure is likely to capture a mixture of at least three factors: effortful attentiveness, interest in the content, and reading comprehension. To the extent that survey experiment participants are low on any or all of these three factors, it will likely bias researchers' treatment effects toward zero, potentially thwarting theoretical innovation in the process. While we cannot confidently speak to which one of these factors MVCs are relatively better or worse at measuring, the larger finding of the present study is that the mock vignette technique offers researchers a simple and effective way of distinguishing those who likely did not attend to the treatment, for one reason or another, from those who did. MVCs therefore enable researchers to conduct hypothesis tests that are more robust to respondent inattentiveness and yet also safeguarded against post-treatment bias.

yielding implausibly large effects, and also differ markedly depending on whether a (factual) manipulation check or timer is used to gauge attentiveness, and differ depending upon which group is designated as the treatment group. See Supplemental Appendix H for details.

REFERENCES

Aarøe, Lene, and Michael Bang Petersen. 2014. "Crowding Out Culture: Scandinavians and Americans Agree on Social Welfare in the Face of Deservingness Cues." *The Journal of Politics* 76 (03): 684–697.

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.

Alvarez, R. Michael, Lonna Rae Atkeson, Ines Levin, and Yimeng Li. 2019. "Paying Attention to Inattentive Survey Respondents." *Political Analysis* 27 (2): 145–62.

Anduiza, Eva, and Carol Galais. 2016. "Answering Without Reading: IMCs and Strong Satisficing in Online Surveys." *International Journal of Public Opinion Research*, May, 1–23.

Ansolabehere, Stephen, Jonathan Rodden, and James M. Jr. Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102 (02): 215–32.

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

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.

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.

———. 2016. "Cheating on Political Knowledge Questions in Online SurveysAn Assessment of the Problem and Solutions." *Public Opinion Quarterly* 80 (4): 858–87.

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

Coppock, Alexander, Thomas J. Leeper, and Kevin Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates across Samples | PNAS." *Proceedings of the National Academy of Sciences* 115 (49): 12441–46.

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

Green, Donald P. 2013. "Breaking Empirical Deadlocks in the Study of Partisanship: An Overview of Experimental Research Strategies." *Politics and Governance* 1 (1): 6-15–15.

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

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments | Political Analysis | Cambridge Core." *Political Analysis* 22 (1): 1–30.

Harden, Jeffrey J., Anand E. Sokhey, and Katherine L. Runge. 2019. "Accounting for Noncompliance in Survey Experiments." *Journal of Experimental Political Science* 6 (3): 199–202.

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.

———. 2016. "Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than Do Subject Pool Participants." *Behavior Research Methods* 48 (1): 400–407.

Hauser, David, Gabriele Paolacci, and Jesse J. Chandler. 2018. "Common Concerns with MTurk as a Participant Pool: Evidence and Solutions," September. https://psyarxiv.com/uq45c/.

Kane, John V., and Jason Barabas. 2019. "No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments." *American Journal of Political Science* 63 (1): 234–49.

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

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. "How Conditioning on Post-Treatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62 (3): 760–75.

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.

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.

Niessen, A. Susan M., Rob R. Meijer, and Jorge N. Tendeiro. 2016. "Detecting Careless Respondents in Web-Based Questionnaires: Which Method to Use?" *Journal of Research in Personality* 63 (August): 1–11.

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 Inference*. Boston, MA, US: Houghton, Mifflin and Company.

Steiner, Peter M., Christiane Atzmüller, and Dan Su. 2016. "Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap." *Journal of Methods and Measurement in the Social Sciences* 7 (2): 52–94.

Thomas, Kyle A., and Scott Clifford. 2017. "Validity and Mechanical Turk: An Assessment of Exclusion Methods and Interactive Experiments." *Computers in Human Behavior* 77 (December): 184–97.

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

Valentino, Nicholas A., Stuart N. Soroka, Shanto Iyengar, Toril Aalberg, Raymond Duch, Marta Fraile, Kyu S. Hahn, et al. 2019. "Economic and Cultural Drivers of Immigrant Support Worldwide." *British Journal of Political Science* 49 (4): 1201–26.

Wood, Dustin, P. D. Harms, Graham H. Lowman, and Justin A. DeSimone. 2017. "Response Speed and Response Consistency as Mutually Validating Indicators of Data Quality in Online Samples." *Social Psychological and Personality Science* 8 (4): 454–64.

Zwaan, Rolf A., Diane Pecher, Gabriele Paolacci, Samantha Bouwmeester, Peter Verkoeijen, Katinka Dijkstra, and René Zeelenberg. 2018. "Participant Nonnaiveté and the Reproducibility of Cognitive Psychology." *Psychonomic Bulletin & Review* 25 (5): 1968–72.