NOTE: THIS ARTICLE WAS CONDITIONALLY ACCEPTED AT POLITICAL SCIENCE RESEARCH & METHODS ON OCTOBER 18, 2022.

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

ABSTRACT

Respondent inattentiveness threatens to undermine causal inferences in survey-based experiments. Unfortunately, existing attention checks may induce bias while diagnosing potential problems. As an alternative, we propose "mock vignette checks" (MVCs), which are objective questions that follow short policy-related passages. Importantly, all subjects view the same vignette before the focal experiment, resulting in a common set of pre-treatment attentiveness measures. Thus, interacting MVCs with treatment indicators permits unbiased hypothesis tests despite substantial inattentiveness. In replications of several experiments with national samples, we find that MVC performance is significantly predictive of stronger treatment effects, and slightly outperforms rival measures of attentiveness, without significantly altering treatment effects. Finally, the MVCs tested here are reliable, interchangeable, and largely uncorrelated with political and socio-demographic variables.

Word Count: 9,000

Keywords: survey experiments; attentiveness; manipulation checks; mock vignettes

Survey experiments have become essential tools for social scientists. And yet, especially given that such research is increasingly being fielded *online* (as opposed to in-person, over the phone, or in a lab), a central concern is that some share of respondents will not be fully attentive. For example, respondents completing surveys remotely may rush through without fully considering what they are reading (Hauser and Schwarz 2016; Thomas and Clifford 2017). This issue presents a critical challenge in experiments: to the extent that a sample is inattentive, "treatments" will not actually be received and, consequently, estimates of treatment effects will likely be biased toward zero.[1] Inattentiveness, therefore, threatens to fundamentally undermine what researchers can learn from their studies, thus stifling theoretical innovation.

Given the seriousness of this challenge, researchers have developed ways of assessing attentiveness in online surveys (e.g., see Druckman 2021). One method comes from Kane and Barabas (2019) who recommend using factual manipulation checks (FMCs) in experiments. Another technique uses question timers to gauge how long respondents spend on a given survey item (Niessen, Meijer, and Tendeiro 2016; Wood et al. 2017). Others employ so-called "screeners", which instruct respondents to provide specific answers to questions wholly unrelated to the experiment (Oppenheimer, Meyvis, and Davidenko 2009). A primary goal of such tools is to help researchers diagnose which respondents are attentive. But, once this individual-level attentiveness is measured, researchers often use such tools for a second purpose—to re-estimate treatment effects for those deemed to be attentive—but this practice raises concerns.

[1] Inattentiveness can thus be understood as measurement error in the independent variable, which leads to "attenuation bias" (Bailey 2021, 145–46).

Specifically, using a post-treatment variable (such as a manipulation check or timer) to remove respondents deemed to be inattentive to the experiment, or to interact with the treatment variable, can introduce covariate imbalances between the randomized treatment and control groups, therein biasing treatment effect estimates (e.g., Aronow, Baron, and Pinson 2019; Coppock 2019; Montgomery, Nyhan, and Torres 2018).

The question then becomes, how can researchers (1) measure individual-level attentiveness to experimental content, 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 technique—*mock vignettes*—for simultaneously accomplishing these three objectives. A mock vignette (MV) contains descriptive information that is substantively similar to content found in political science experimental 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 attentiveness to the MV. We refer to these items as *mock vignette checks* (MVCs). From the respondent's perspective, 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 are asked follow-up questions (e.g., Mutz 2011; 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 perform better on the MVCs—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 conditional average treatment effect (CATE) by interacting the treatment with performance on the MVCs. To the extent that inattention is downwardly biasing a

treatment effect, then, the researcher should observe *stronger* treatment effects when analyzing those who perform better (versus worse) on the MVCs. Most importantly, by virtue of appearing *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 common 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 ITT 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. We also investigate whether there exist 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 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. Finally, we find that MVCs perform modestly better, on several dimensions, than a common alternative method for assessing respondent attentiveness.

Mock vignettes thus enable researchers to test hypotheses on respondents who are likely to have been attentive to their experiment. Further, MVs can be used in 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, 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, measures of complexity, and item response theory (IRT) analyses.

Though designed to resemble vignette-based experiments, MVs/MVCs can be of potential value in other related experiments (e.g., conjoint, question-wording, or list experiments). Our findings suggest that by implementing a mock vignette before their experiment, researchers are better equipped to overcome the challenge of respondent inattentiveness, and can therefore perform fairer, more reliable, and more robust tests of their hypotheses.

INATTENTIVENESS & POST-TREATMENT BIAS

Whether researchers attempt to identify them or not, experiments will likely contain a sizable share of inattentive respondents. Respondents may 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 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.[2] Imprecise estimates, and/or null or substantively weak effects may be mistakenly interpreted as a flawed theory or design, rather than as a consequence of respondent noncompliance.

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 more 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 ostensibly banal questions about unrelated topics that discreetly ask respondents to answer in a specific fashion (Berinsky, Margolis, and Sances 2014; Hauser and Schwarz 2015; Oppenheimer, Meyvis, and Davidenko 2009). In these studies and elsewhere, answering such manipulation check questions correctly (incorrectly) is indicative of greater (less) respondent attentiveness. An alternative approach involves the use of question timers, wherein the amount of time that respondents spend on a given screen (e.g., an experimental vignette) is recorded. For such time measures (or, latencies), low scores indicate insufficient attention (Harden, Sokhey, and Runge 2019, 3; Niessen, Meijer, and Tendeiro 2016; Wood et al. 2017; Zwaan et al. 2018).

[2] 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.

What can be done with these measures? On one hand, they can be used to gauge the overall share of attentive respondents participating in the study as a whole (and, in the case of FMCs and timers, also the share of respondents attentive to a particular experimental condition). This serves as a useful diagnostic tool to help adjudicate between competing interpretations of a given result (e.g., a non-significant result being due to a misguided theory and/or hypothesis versus being due to substantial respondent inattentiveness). FMCs also have the added benefit of enabling the researcher to ensure that responses correlate with treatment assignment, thereby functioning not only as a measure of attention to the content but also as evidence that the manipulation itself was sufficiently perceived.

However, researchers have also tended to use such measures when 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; Druckman 2021). Along similar lines, 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*, potentially yielding a biased estimate of the treatment effect. Worse still, researchers have limited statistical ability to completely rule out the possibility of post-treatment bias (Acharya, Blackwell, and Sen 2016; Montgomery, Nyhan, and Torres 2018, 772–73).

With these interrelated challenges in mind, we propose an alternative technique for measuring respondent attentiveness to experimental content that can be easily incorporated into analyses of survey experiments (including more elaborate experiments, such as factorial designs and conjoint experiments). 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 a post-treatment measure. Manipulation checks and timers on experimental items, therefore, risk introducing bias when employed in the estimation of treatment effects. Thus, while such a measure is ideal because it directly gauges attentiveness to one's experiment, an alternative is needed if we wish to also re-estimate treatment effects on the attentive respondents.

In proposing such a measure, we first reason that, because respondent attentiveness varies throughout the course of a survey (e.g., Alvarez et al. 2019), the measure should be as close in proximity to the experiment as possible—ideally, immediately pre-treatment. Second, 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*. Designed as such, a respondent's attentiveness to this pre-treatment content can thus function as a proxy for the respondent's attentiveness to the actual experiment's vignettes and outcome measure(s).

We therefore propose that researchers employ 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, display information to respondents (Steiner, Atzmüller, and Su 2016). 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 and MVCs.

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. 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). When *multiple* MVCs are employed (see examples below) an attentiveness scale can be constructed as one would for other social science concepts. Following the MV and MVC(s), each respondent is then randomly assigned to an experimental condition.

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 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 while bypassing the threat of post-treatment bias (Montgomery, Nyhan, and Torres 2018, 770–71). Similarly, the researcher can test the robustness of their ITT estimate by interacting the treatment indicator with MVC performance: if a treatment were 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 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 better MVC performance is associated with stronger treatment effects. The following sections directly investigate this hypothesis.

DATA & METHODS

We conducted five studies, beginning in May of 2019 through February of 2020, featuring U.S. adults. 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, which employed quotas to obtain a sample nationally representative in terms of age, race/ethnicity, and geographic region. Lastly, and recruited by the National Opinion Research Center (NORC), the NORC study features a nationally-representative probability sample from NORC's "AmeriSpeak Omnibus" survey.

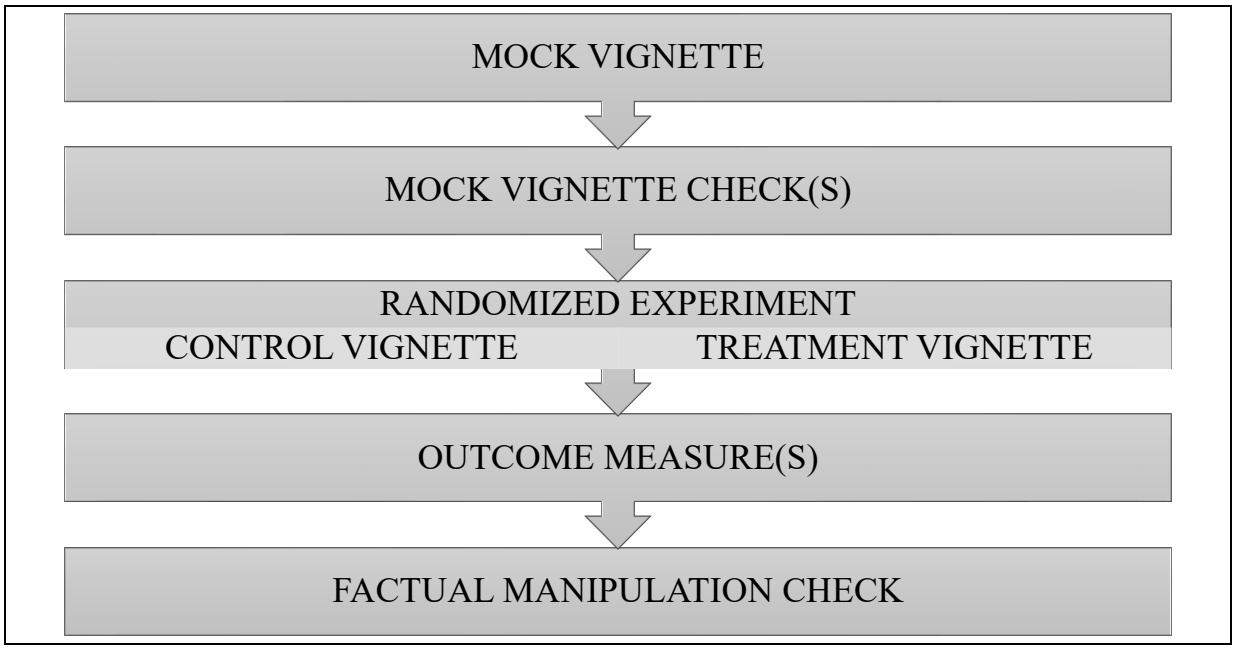
TABLE 1. Overview of Samples, Mock Vignettes, and Experiments

	MTurk 1 (n=603)	**Qualtrics** (n=1,040)	**NORC** (n=1,021)	**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). N sizes reflect sample used for replicating experiment; Qualtrics and NORC studies had 25% of sample randomly assigned to not receive an MV/MVCs.

Each of these studies featured the same basic design depicted in Figure 1. Respondents in each study were shown the same mock vignette (MV), then answered at least one factual question aimed at checking attentiveness to this MV (with no ability to "go back" to the 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. Timers were used on each screen to record the amount of time (in milliseconds) respondents spent on each screen. 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 (unseen by respondents). These latter steps permit benchmarking MVCs versus other measures of attention to the experiment itself (discussed in detail below).

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). Total N = 5,644. Samples sizes in first round: MV (1) 1,112; (2) 1,116; (3) 1,109; (4) 1,128; Experiment (1) 1,411; (2) 1,411; (3) 1,410; (4) 1,412. Sample sizes in second round: MV (1) 1,355; (2) 1,360; (3) 1,357; (4) 1,363; Experiment (1) 1,358; (2) 1,362; (3) 1,359; (4) 1,356.

In the fifth experiment—the Lucid 1 study (n=6,216)—each respondent experienced two iterations of the design featured in Figure 1.[3] In other words, within each of two separate rounds, each respondent was randomly assigned one of four possible MVs (and its corresponding MVCs), as well as one 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 thus 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 1 study.

[3] A second Lucid study is detailed below.

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. 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 "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, as well as additional details regarding source material, can be found in the Supporting Information (SI) file (section A).

Except for the "Minimum Sentencing" MV (which only featured 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, responses to each MVC are coded as either incorrect (0) or correct (1). In every study except NORC, respondents were required to offer

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.

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.[4] 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).

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 support for the demonstration to continue (again, measured on a 7-point scale ranging from *strongly oppose* to *strongly support*). This experiment has also been replicated in prior studies (e.g., Mullinix et al. 2015).

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

[4] The nature of the questions and response options was kept as similar as possible across MVs. Generally, the first MVC measures attentiveness to the general topic, while the second and third MVCs measure attentiveness to the first half and second half of the MV, respectively.

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" study 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. The text for all aspects of the replicated studies—i.e., the vignettes, outcome response options, and factual manipulation checks—can be found in the SI (section 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 the Qualtrics study had a

passage rate of 55%.[5] For the NORC and MTurk 2 studies, which featured one MV with three MVCs, passage rates for any given MVC ranged from 36% to 81%, and 44% to 80%, respectively. In the Lucid 1 study, passage rates were generally between 50% and 80%.

MVC Performance and Treatment Effect Size

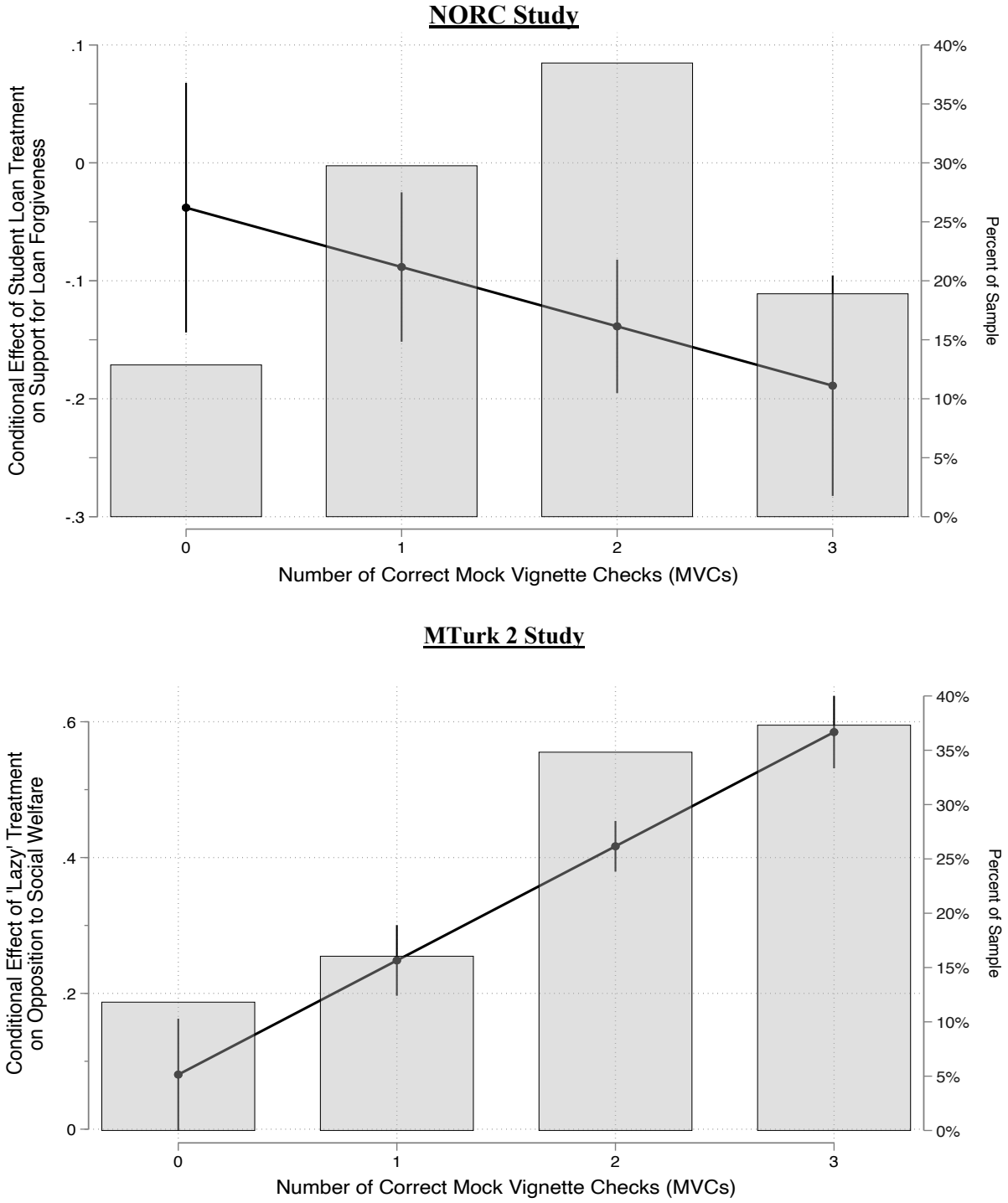
We now investigate whether MVC passage is associated with larger treatment effect sizes. Given word-limit constraints on this manuscript, and because the MTurk 1 and Qualtrics studies were unusual in that they only featured one MVC, we detail the results for these two studies in the SI (section C). In brief, for both studies, we find 1) stronger treatment effects among MVC passers relative to non-passers, 2) that treatment effects among MVC passers were statistically significant at the $p<.01$ level, and 3) that treatment effects among MVC non-passers failed to attain significance at conventional levels. This serves as preliminary evidence that MVCs identify respondents for whom experimental treatments will be more efficacious and, conversely, identify respondents who are less attentive and, thus, less affected by the treatment information.

Compared to the MTurk 1 and Qualtrics studies, a major advantage of the NORC and MTurk 2 studies is that, while each features only one MV, there are *three* accompanying MVCs. Employing multiple MVCs yields a scaled measure of attentiveness that is likely to contain less measurement error than that of a single MVC.

Figure 2 displays the conditional average treatment effect (CATE), in both the NORC (top panel) and MTurk 2 (bottom panel) studies, across MVC performance. In each study, the dependent variable has been rescaled to range from 0 to 1 to enhance interpretability. Histograms are also

[5] Observing a relatively higher level of attentiveness in the MTurk samples is consistent with research by Hauser and Schwarz (2016).

FIGURE 2. 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). Top (bottom) panel shows that the negative (positive) effect observed in original experiment grows larger in magnitude as MVC performance increases. Histogram represents the percent of the sample correctly answering *x* MVCs. Total N=744 (NORC) and 804 (MTurk Study 2).

featured to convey the distribution of MVC performance within each study, with the right *y*-axis of each graph displaying the percentage of the sample passing a given number of MVCs.

The top panel of Figure 2 (NORC data) features the "Student Loan" experiment, in which the treatment is designed to significantly *reduce* support for student loan forgiveness (Mullinix et al. 2015). We observe 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.[6] This interaction between treatment and MVC 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 percentage-point effect that was observed among those who did not answer any MVCs correctly (approximately 13% of the sample). As this analysis 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.[7]

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

[7] As a further illustration, among the 140 respondents who passed all 3 MVCs, the effect was .23 ($p < .01$). Post-hoc calculations confirm that power is high (power= .96, two-tailed, alpha=.05). Conversely, given the smaller effect among those who passed 0 MVCs (effect=.076, se=.078), to have a well powered study with an effect of this size would require several times as many experimental subjects.

The results for the MTurk 2 study (see bottom panel of Figure 2) are even more pronounced. Replicating the social welfare deservingness experiment (Aarøe and Petersen 2014), the bottom panel of Figure 2 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.[8]

Because this latter set of analyses involved an MVC scale rather than a single binary measure (which researchers could simply use to subset their data), these results exemplify how MVCs can be easily incorporated into analyses: 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, while still avoiding 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 test of one's hypothesis. In addition, this approach is potentially valuable 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.

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

We now turn to the Lucid 1 study, 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 performance. Using additional models, we also 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.[9] 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 an indicator of treatment status and MVC represents the respondent's score on the MVC scale (i.e., the number of correct MVCs).[10]

[9] Fixed effects and random intercept models were also estimated. However, this yielded substantively identical results.

[10] We separately assess the robustness of the linearity assumption underlying this interaction, and find that the data are consistent with a linear multiplicative model (see Appendix K).

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

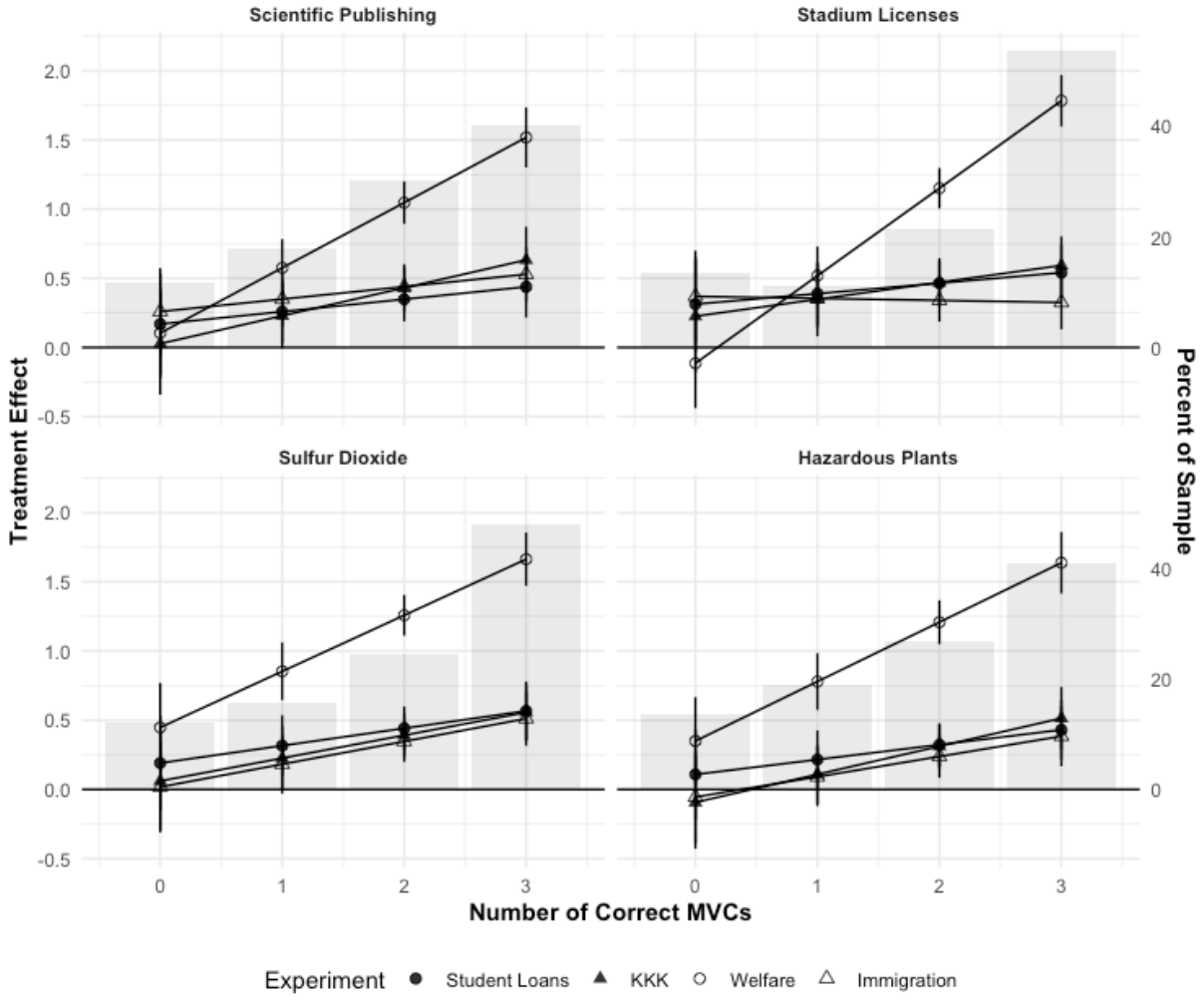
Notes: Lucid study. OLS regression coefficients with standard errors clustered by respondent. Outcome is standardized within each experiment (control group standard deviations). Mock Vignette Check Score ranges from 0 to 3. ***p<0.001 (one-tailed hypothesis tests).

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.[11] 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.

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 3. This figure

[11] Given the need to aggregate across multiple studies with different outcome measures, we standardize our outcomes using control group standard deviations (SDs). However, three of the four experiments feature seven-point Likert scales with SDs approximately equal to 2, and thus, we also report raw scale quantities to facilitate the interpretation of effects. Though the immigration study did not use a single seven-point Likert scale, outcomes were measured using three items that sum to a score of seven. The SD for this study is 1.82.

FIGURE 3: 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.

reveals that MVC performance is positively associated with CATEs in 15 out of the 16 MV-Experiment combinations.[12] The relationship between MVC performance and CATEs is strongest in

[12] CATE estimates among those assigned to the stadium licenses-immigration pair slightly decrease as a function of correct MVCs, though not statistically significantly so.

the *Welfare* experiment. The latter's ITT 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 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 3 also suggests that the relationship between MVC performance and CATEs is relatively similar regardless of the particular MV that is used.[13]

Validating MVC Scores with Other Measures of Attentiveness

An implication of the aforementioned findings is that MVC performance should be associated with better performance on *other* measures of attentiveness to the survey experiment. 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 1 study MVCs had pairwise correlations ranging from .55 to .63 ($p<.001$), and Cronbach's alpha (α) values ranging

[13] We conducted an explicit test of this possibility, and find differences between MVs—in terms of predicting larger CATEs—to be minimal and not statistically discernible from zero (see SI (section F)).

from.60 to .74.[14] Further, in the Lucid 1 study, the grand pairwise correlation between round 1 and round 2 MVC performance (i.e., between the two MVC scales) was quite strong at .60 ($p<.001$).[15]

We also investigate correlations with question timers, for which 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 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, 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.

Due to spatial limitations, we present the full results of our analyses in the SI (section E). To summarize results for the MTurk, Qualtrics and NORC studies, better performance on the MVC consistently predicts greater latency (i.e., more time spent) on (1) the mock vignette itself, (2) the experimental vignettes, and (3) the experiment's outcome measure. These differences were positive in

[14] The "Same-Day Registration" MVCs displayed 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)).

[15] While our primary interest in this section is the relationship between MVC performance and attentiveness to *experimental* content, it is notable that these correlations are considerably higher than those found for a variety of other kinds of attentiveness measures (e.g., see Niessen, Meijer, and Tendeiro 2016), including instructional manipulation checks (IMCs, also known as "screeners"; (Berinsky, Margolis, and Sances 2014; Thomas and Clifford 2017)).

sign and statistically significant at $p<.05$ or below in all but once instance.[16] For example, in the KKK experiment (Qualtrics study), 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.) In every Lucid 1 experiment, better MVC performance predicts significantly greater latencies. Thus, in 39 out of 40 separate tests, we find that better MVC performance is associated with significantly more attentiveness to experimental content. Further, in every case, those who passed the MVC spent significantly more time on the survey itself.[17]

Lastly, we analyze responses to factual manipulation checks (FMCs), which also aim to measure attentiveness to the actual experiment's vignettes. We find a remarkably strong relationship between MVC performance and FMC passage: 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 1 study, these effects were even stronger, ranging from 41 to 68 percentage points (see SI (section E) for details). Thus, in 8 out of 8 separate tests, we find that better MVC performance predicts a significantly greater likelihood of correctly answering a factual question about the contents of the experiment. This serves as further evidence that MVCs can, as

[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 the sample, all other analyses with Qualtrics data exclude these "speeders".

intended, function as a pre-treatment, proxy measure of the extent to which respondents are attentive to one's 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, consequently, weaker treatment effects.

To investigate these potential concerns, we designed the Qualtrics, NORC and Lucid 1 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 (i.e., a replication of the original study without a mock vignette, nor the mock vignette checks, for comparison purposes). Across these studies, treatment effects were substantively and statistically similar regardless of whether an MV was featured. We also find no significant differences in the *variance* of the outcome measures depending upon whether a MV was or was not viewed, suggesting that MVs also do not induce heterogeneity in responses to the outcome measure. Due to limited space, we feature the full results of these analyses in the SI (section G).

Demographic Patterns in MVC Performance

A common issue with analyzing attentive respondents is that the subset of attentive respondents may differ from one's entire sample on a variety of demographic and politically-relevant variables (e.g., Thomas and Clifford 2017). Indeed, such differences would be expected insofar as attentiveness is not randomly distributed in the population. In each of our studies, we thus 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 space constraints, the full set of results can be found in the SI (section D). 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 prior 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 or 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 (e.g., political interest).* This latter finding helps assuage the potential concern that, for example, only highly educated and/or politically interested respondents will be able to correctly answer MVCs.

Nevertheless, as prior studies duly note (e.g., Thomas and Clifford 2017), we caution that analyzing the attentive may alter the demographic composition of the sample. This may be important when researchers desire *descriptive* results that can apply to the broader population, and/or when such demographic variables significantly moderate a particular experiment's effect. Researchers can increase transparency 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.[18]

Lastly, 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

[18] Researchers can also 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 1 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.

if the attentive sub-sample differs demographically from the full sample, this potentially poses constraints on the external validity of the results, rather than threatening the survey experiment's internal validity. Moreover, 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). In sum, while the external validity of one's findings remains an important consideration, neglecting to account for inattentiveness whatsoever risks obtaining treatment effect estimates that are downwardly biased, potentially yielding null findings and undermining one's study.

Detecting Significant Effects Among Attentive Sub-Samples

Analyzing a *subset* of one's sample raises practical questions concerning statistical power and, specifically, whether one can still detect statistically significant treatment effects when analyzing the attentive sub-group. We investigate these concerns in each of our studies. 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 (see SI (section H) for details).

Comparison With Instructional Manipulation Checks

While we emphasize that MVCs can be used in conjunction with other kinds of attentiveness measures, we fielded a separate study via Lucid in 2021 (total n=9,000; "Lucid 2") to investigate how MVCs performed relative to an existing method for assessing inattentiveness in survey experiments—i.e., instructional manipulation checks (IMCs), or "screeners" (Berinsky, Margolis and Sances 2014).

We document the details of this investigation in SI (section I). Overall we find that MVCs slightly outperform IMCs on several dimensions. In particular, the MVC scale tended to yield slightly larger CATEs than the IMC scale, which is consistent with our argument that MVCs, by design, should be stronger predictors of attentiveness to the vignette in one's experiment. Echoing this latter point, we also find that MVC performance predicts significantly longer time spent on experimental stimuli and outcome measures relative to IMC performance (though their respective effects on total survey duration were nearly identical), as well as significantly better performance answering post-outcome FMCs (approximately 8 percentage points, $p<.05$). Lastly, with the exception of age (which, though significantly associated with both MVC and IMC performance, is more strongly associated with the former), demographic and political variables operate remarkably similarly in predicting MVC vs. IMC performance.[19]

DISCUSSION & CONCLUSION

The growth of experimental social science has exploded in recent years due to technological advances that allow survey experiments to be programmed and fielded online with relative ease (e.g., Druckman 2021). Yet, a persistent challenge arising from this method is 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.

[19] Further, we reanalyze the Lucid 1 data using the 2SLS approach suggested by Harden, Sokhey and Runge (2019). Overall, we find the implementation of this method to be more complex than the MV approach we propose. In particular, there are far more modeling choices the researcher must make, and these decisions lead to a wide range of substantively different results (see SI (section J) for details).

Taken together, we believe our findings indicate that survey researchers can benefit substantially from featuring MVs and MVCs in their studies, and with few downsides beyond the need to include additional items in their surveys.[20] In fact, we found the inclusion of MVs to be beneficial despite using survey firms that feature pre-screened opt-in samples and/or flag and remove inattentive respondents before data collection concludes.[21]

In our SI, we provide text and performance analytics for all of MVs and MVCs used in this study. If scholars wish to use these items, or construct their own, we emphasize the following suggestions based upon our studies' designs (see also Table A7 in the SI). 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, we recommend that scholars present MVCs as forced response questions to avoid missing data, and with

[20] By increasing the survey length, including an MV and MVC(s) may in some cases increase the financial cost of fielding a survey experiment. Researchers thus need to consider whether accounting for inattentiveness in their study is worth this additional cost. Alternatively, MVCs could potentially be used to "screen out" inattentive respondents, though researchers would need to ensure that this does not create difficulties for generalizing results to the underlying population of interest, nor would it be likely to completely eliminate inattentiveness to one's experiment.

[21] In the case of Qualtrics, for example, respondents whose total time is below 2 standard deviations of the mean completion time are automatically "screened out"; for details, see https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/response-quality/).

the "back button" disabled 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 (as featured in most of our studies) are advisable where possible. Fourth, block-randomizing based upon responses to at least one MVC would help ensure that attentiveness is balanced across experimental conditions (e.g., Gerber and Green (2012) find modest benefits of this practice in small samples).[22] Finally, we urge researchers to be fully transparent by presenting the ITT for the sample as a whole before presenting re-estimated treatment effects on those deemed to be attentive and/or presenting whether (and to what degree) the estimated treatment effect increases in magnitude at higher levels of attentiveness.

Insofar as it gauges pre-treatment attentiveness to vignette-based content, our findings indicate the MV approach comes with potential advantages over alternative approaches, though we emphasize that these various techniques need not be treated as mutually exclusive. For instance, MVCs could be used in conjunction with timers, instructional manipulation checks (IMCs), and related techniques to assess general attentiveness (e.g., Oppenheimer, Meyvis, and Davidenko 2009; Vraga, Bode, and Troller-Renfree 2016). If the various measures scale together sufficiently well, they could be combined into a single continuous measure of attentiveness; otherwise, researchers may separately report CATE estimates, for example, using each measure of attentiveness that was employed.

In addition, though we identify several distinct advantages to the MV approach, its use does not obviate the need for other tools that gauge attentiveness to experimental content, such as manipulation checks. Treatment-relevant factual manipulation checks (FMC-TRs), for example,

[22] For example, we used block randomization in the MTurk 2 study based upon whether the respondent correctly answered the first MVC (about the general topic of the MV).

provide crucial information about the degree to which experimental manipulations were actually perceived, while more conventional manipulation checks (i.e., subjective manipulation checks) help researchers determine whether the experimental manipulation is affecting the theorized independent variable of interest. By including such items, the researcher is far better able to gauge the extent to which either respondent inattentiveness to experimental content and/or an ineffective manipulation, respectively, are influencing the results of the experiment.

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 also presents a useful avenue to explore in future research.

In sum, the mock vignette technique offers researchers a simple and effective way of distinguishing those who likely did not attend to their survey experiments, for one reason or another, from those who did. In so doing, MVCs enable researchers to conduct hypothesis tests that are more robust to respondent inattentiveness and also avoid post-treatment bias. We believe this technique will therefore equip researchers with an ability to understand their results at a deeper level than what the simple ITT estimate permits, and thus allow them to learn more from their experimental studies.

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): 684–97.

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.

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.

Bailey, Michael A. 2021. *Real Stats: Using Econometrics for Political Science and Public Policy*. 1st ed. New York, NY: Oxford University Press.

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.

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.

Druckman, James N. 2021. *Experimental Thinking: A Primer on Social Science Experiments*. New York, NY: Cambridge University Press. https://faculty.wcas.northwestern.edu/~jnd260/pub/Druckman%20Experimental%20Thinking%20Fall%202020%20Submitted.pdf.

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

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

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): 109–38.

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.

Vraga, Emily, Leticia Bode, and Sonya Troller-Renfree. 2016. "Beyond Self-Reports: Using Eye Tracking to Measure Topic and Style Differences in Attention to Social Media Content." *Communication Methods and Measures* 10 (2–3): 149–64.

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.

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

SUPPORTING INFORMATION (SI)

TABLE OF CONTENTS

APPENDIX A:
MOCK VIGNETTE TEXT, ANALYTICS, & PROTOCOLS

Mock Vignette 1: "Same-Day Registration" (based upon information from the National Conference of State Legislatures).

Mock Vignette

Many state legislatures are currently considering enacting "same-day registration" policies, which would allow residents of the state (who are eligible to vote) to register and vote within the same day.

As with any policy, there are many factors to consider, including factors regarding potential costs of implementation. According to the National Conference of State Legislatures, "Same day registration procedures vary within states, and so costs vary as well. Some states indicate there is little to no additional cost in implementing same day registration, especially those that have had this option available for a long time. Some costs that may be associated with implementing same day registration include increased election staff or poll workers to process same day registrations. This extra administrative task can be time consuming at the same day registration site and when verifying registration information after the election. Many states report this is more a reallocation of costs and resources, though, rather than an additional cost."

Mock Vignette Check 1 [new screen; randomized response options; correct answer shaded]

Which policy area was discussed in the article you just read?

- Voter identification policies
- Voting age policies
- Voter registration policies
- Voting privacy policies
- Voting location policies

Mock Vignette Check 2 [new screen; randomized response options; correct answer shaded]

In the article you just read, which specific organization was quoted regarding "same-day registration" policies?

- State Board of Elections
- Council on Foreign Relations
- National Conference of State Legislatures
- Bureau of Legislative and Electoral Processes
- National Governors Group

Mock Vignette Check 3 [new screen; randomized response options; correct answer shaded]

- According to the article you just read, many states report that:
- Same-day registration has "resulted in many more voters coming to the polls"
- There is now "significantly greater gubernatorial oversight of the voting process"
- "Increased election staff" may be one cost of implementation
- Policymakers "do not believe this policy will significantly change voter turnout"
- They have "recently reversed their position on voting registration policies"

Table A1. Same-Day Registration Mock Vignette and MVC Analytics

Mock Vignette (MV)	
Sample(s) Used	NORC
Word Count	158
Average Time Spent on Screen [95% CI]	38.13 seconds [34, 43]
"Flesch Reading Ease" Score	29
Mock Vignette Check 1	
Proportion Answering Correctly [95% CI]	.81 [.78, .83]
Difference versus Random Guessing (p-value)	.61 (<.001)
Mock Vignette Check 2	
Proportion Answering Correctly [95% CI]	.36 [.32, .40]
Difference versus Random Guessing (p-value)	.16 (<.001)
Mock Vignette Check 3	
Proportion Answering Correctly [95% CI]	.47 [.43, .50]
Difference versus Random Guessing (p-value)	.27 (<.001)

Notes: Flesch Reading Ease score obtained from https://datayze.com/readability-analyzer. Scores between 50 and 70 are approximately equivalent to high-school English; below 30 is college graduate-level English. MVC "Proportion Answering Correctly" reports logistic 95% confidence intervals (CI). "Difference versus Random Guessing (p-value)" based on results from a one-sample difference-in-proportions test, wherein the probability of randomly guessing the correct response equals 1/5 (.20). Two-tailed p-value reported.

Mock Vignette 2: "Scientific Publishing" (original source material)

Intro Screen

Next, we would like to ask you an additional question on a different topic. Please read the following excerpt from a recent magazine article.

Mock Vignette

A Passage from a Recent Magazine Article:
More than 125 scientific societies and journal publishers are urgently warning lawmakers not to move forward with a rumored 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 the scientific enterprise. Lawmakers won't comment on whether they are considering a policy that would change publishing rules, and society officials say they have learned no details. But if the rumor is accurate, the order would represent a major change from current U.S. policy, which allows publishers to withhold federally-funded research from the general public for up to 1 year.

Mock Vignette Check 1 [new screen; randomized response options; correct answer shaded]

What was the topic of the magazine excerpt you just read?

- Literary magazines
- Scientific research publishing
- Arts funding
- English education
- Immigration policy
- Funding for space exploration

Mock Vignette Check 2 [new screen; randomized response options; correct answer shaded]

Regarding the rumored change in policy that was discussed, the magazine excerpt indicated that:

- Lawmakers won't comment on whether they are considering it
- Legal scholars stated the change in policy would be challenged in courts
- Journal publishers have already begun preparing for the change in policy
- Scientific researchers are divided in terms of their support for the policy
- All of the above
- None of the above

Mock Vignette Check 3 [new screen; randomized response options; correct answer shaded]

According to the magazine excerpt you just read, current policy allows federally-funded research to be withheld from the general public for up to:

- 1 month
- 6 months
- 1 year
- 3 years
- 5 years
- None of the above

Table A2. Scientific Publishing Mock Vignette and MVC Analytics

Mock Vignette (MV)	
Sample(s) Used	1. MTurk (Study 2) 2. Lucid
Word Count	128
Average Time Spent on Screen [95% CI]	**MTurk:** 37.70 seconds [33, 42] **Lucid:** 56.93 seconds [51, 63]
"Flesch Reading Ease" Score	43
Mock Vignette Check 1	
Proportion Answering Correctly [95% CI]	**MTurk:** .80 [.77, .83] **Lucid:** .78 [.76, .80]
Difference versus Random Guessing (p-value)	**MTurk:** .64 (<.001) **Lucid:** .61 (<.001)
Mock Vignette Check 2	
Proportion Answering Correctly [95% CI]	**MTurk:** .44 [.41, .48] **Lucid:** .50 [.48, .52]
Difference versus Random Guessing (p-value)	**MTurk:** .27 (<.001) **Lucid:** .33 (< .001)
Mock Vignette Check 3	
Proportion Answering Correctly [95% CI]	**MTurk:** .73 [.70, .76] **Lucid:** .71 [.69, .73]
Difference versus Random Guessing (p-value)	**MTurk:** .56 (<.001) **Lucid:** .54 (<.001)

Notes: Flesch Reading Ease score obtained from https://datayze.com/readability-analyzer. Scores between 50 and 70 are approximately equivalent to high-school English; below 30 is college graduate-level English. MVC "Proportion Answering Correctly" reports logistic 95% confidence intervals (CI). "Difference versus Random Guessing (p-value)" based on results from a one-sample difference-in-proportions test, wherein the probability of randomly guessing the correct response equals 1/6 (.1667). Two-tailed p-value reported.

Mock Vignette 3: "Stadium Licenses" (original source material)

Intro Screen

Next, we would like to ask you additional questions on a different topic. Please read the following passage from a recent magazine article.

Mock Vignette

A Passage from a Recent Magazine Article:

Officials in a midsize town have been working for four years on a plan to produce an event license to cover all of the major events that occur at the town's local stadium, which hosts concerts and home sports games. The application would be submitted each January and list all events expected to occur at the stadium over the next 12 months. If an unlisted event emerges during the year, lawmakers could hold a special hearing on the event, or accept it without a hearing and add it into the existing license. To assist with this plan, lawmakers filed legislation that would change state licensing laws so that annual event licenses will expire within one year. "This makes a minor change to current law, which provides that all licenses issued shall expire on December 31 of each year," a lawmaker said.

Mock Vignette Check 1 [new screen; randomized response options; correct answer shaded]

What was the topic of the magazine article you just read about?

- Stadium funding
- Event licensing
- Political polarization
- City budgeting
- Election monitoring policy
- Campaign finance reform

Mock Vignette Check 2 [new screen; randomized response options; correct answer shaded]

What, according to the magazine article, is the ultimate goal of the policy?

- Property taxes will finance the construction of a new stadium
- A single license will cover all events occurring in a stadium
- Construction companies will be adequately compensated
- The town will attract a more diverse workforce
- Local residents will receive discounted rates
- The town will begin researching noise control technologies

Mock Vignette Check 3 [new screen; randomized response options; correct answer shaded]

After the new event license law goes into effect, what, according to the magazine article, could happen if an unlisted or unplanned event emerges during the year?

- The event organizers must pay a small surcharge to the Town Board
- There may be a special hearing held by lawmakers
- Money will be diverted from a special fund
- Team managers will decide whether the event takes place or not
- All of the above
- None of the above

Table A3. Stadium Licenses Mock Vignette and MVC Analytics

Mock Vignette (MV)	
Sample(s) Used	Lucid
Word Count	148
Average Time Spent on Screen [95% CI]	63.31 seconds [47, 79]
"Flesch Reading Ease" Score	50
Mock Vignette Check 1	
Proportion Answering Correctly [95% CI]	.74 [.72, .76]
Difference versus Random Guessing (p-value)	**.57 (<.001)**
Mock Vignette Check 2	
Proportion Answering Correctly [95% CI]	**.79 [.77, .81]**
Difference versus Random Guessing (p-value)	**.62 (<.001)**
Mock Vignette Check 3	
Proportion Answering Correctly [95% CI]	**.62 [.60, .64]**
Difference versus Random Guessing (p-value)	**.45 (<.001)**

Notes: Flesch Reading Ease score obtained from https://datayze.com/readability-analyzer. Scores between 50 and 70 are approximately equivalent to high-school English; below 30 is college graduate-level English. MVC "Proportion Answering Correctly" reports logistic 95% confidence intervals (CI). "Difference versus Random Guessing (p-value)" based on results from a one-sample difference-in-proportions test, wherein the probability of randomly guessing the correct response equals 1/6 (.1667). Two-tailed p-value reported.

Mock Vignette 4: "Sulfur Reductions" (original source material)

Intro Screen
Next, we would like to ask you additional questions on a different topic. Please read the following passage from a recent magazine article.

Mock Vignette

A Passage from a Recent Magazine Article:

The International Maritime Organization (IMO), the industry's regulator, will require all ships to cut the level of sulfur in their engine emissions beginning January 1st. The limit reduces the sulfur dioxide (SO2) that ships emit into the atmosphere via the ship's funnel. Therefore, policymakers expect that there will be a reduction in the SO2 that finds its way into the air. It may seem like a small change, but the effects will ripple across the oil value chain. For example, many ships will comply by investing in scrubbers that strip the sulfur out of the exhaust. But, there is a lot of worry over the possibility that ships will divert air pollutants directly into the sea, leading to greater pollution in the ocean. The other issue is that the regulation does not currently require refiners to remove the sulfur at its origin.

Mock Vignette Check 1 [new screen; randomized response options; correct answer shaded]

What was the topic of the magazine article you just read about?

- Industrial chemical solutions
- New steel tariffs
- Sulfur reductions
- Plane cargo limits
- Air travel regulations
- Fishing licensing reform

Mock Vignette Check 2 [new screen; randomized response options; correct answer shaded]

Which organization, according to the magazine article, was responsible for the rule change?

- National Science Foundation
- International Maritime Organization
- International Monetary Fund
- Industrial Manufacturing Organization
- National Oceanic and Atmospheric Association
- Government Accountability Office

Mock Vignette Check 3 [new screen; randomized response options; correct answer shaded]

What, according to the magazine article, is a possible consequence of the rule change being discussed?

- Increased pollution into the ocean
- Decreased profits for businesses
- Increased corruption in government
- Increased health risks of citizens
- All of the above
- None of the above

TABLE A4. "Sulfur Reductions" Mock Vignette and MVC Analytics

Mock Vignette (MV)	
Sample(s) Used	Lucid
Word Count	149
Average Time Spent on Screen [95% CI]	52.82 seconds [48, 58]
"Flesch Reading Ease" Score	49
Mock Vignette Check 1	
Proportion Answering Correctly [95% CI]	.80 [.78, .82]
Difference versus Random Guessing (p-value)	.63 (<.001)
Mock Vignette Check 2	
Proportion Answering Correctly [95% CI]	.61 [.59, .63]
Difference versus Random Guessing (p-value)	.44 (<.001)
Mock Vignette Check 3	
Proportion Answering Correctly [95% CI]	.67 [.65, .69]
Difference versus Random Guessing (p-value)	.50 (<.001)

Notes: Flesch Reading Ease score obtained from https://datayze.com/readability-analyzer. Scores between 50 and 70 are approximately equivalent to high-school English; below 30 is college graduate-level English. MVC "Proportion Answering Correctly" reports logistic 95% confidence intervals (CI). "Difference versus Random Guessing (p-value)" based on results from a one-sample difference-in-proportions test, wherein the probability of randomly guessing the correct response equals 1/6 (.1667). Two-tailed p-value reported.

Mock Vignette 5: Hazardous Plants (original source material)

Intro Screen

Next, we would like to ask you additional questions on a different topic. Please read the following passage from a recent magazine article.

Mock Vignette

A Passage from a Recent Magazine Article:
A new law regarding hazardous vegetation (such as trees, bushes, plants, etc.) has been in effect since early this year after being passed by the local city council. The law, which requires brush-clearing for properties, gives the county the power to hire contractors to remove hazardous vegetation if property owners do not comply with the law, and then charge property owners for the work done. This legislation eventually received total support from the board of supervisors. However, the original law had been changed after residents criticized a first draft of the proposal, citing concern over potentially massive fines for property owners and other issues. But the late adoption of the law complicated the timing of its enforcement, creating uncertainty over how forcefully to push property owners to clear their brush when summer temperatures, and fire danger, are high.

Mock Vignette Check 1 [new screen; randomized response options]

What was the topic of the magazine article you just read?

- Insect traps
- Forest protection
- Hazardous vegetation
- Climate change
- Sewage routing
- Property taxes

Mock Vignette Check 2 [new screen; randomized response options]

Who, according to the magazine article, was responsible for passing the policy change?

- Environmental Protection Agency
- Water and Sewer Department
- Local city council
- Federal government
- Trump administration
- State legislature

Mock Vignette Check 3 [new screen; randomized response options]

What, according to the magazine article, was a criticism of the policy when it was originally proposed?

- Potentially massive fines
- Increased tax rates
- Warmer temperatures

- Increased health risks
- All of the above
- None of the above

TABLE A5. "Hazardous Plants" Mock Vignette and MVC Analytics

Mock Vignette (MV)	
Sample(s) Used	Lucid
Word Count	145
Average Time Spent on Screen [95% CI]	67.01 seconds [52, 82]
"Flesch Reading Ease" Score	40
Mock Vignette Check 1	
Proportion Answering Correctly [95% CI]	.81 [.80, .83]
Difference versus Random Guessing (p-value)	.64 (<.001)
Mock Vignette Check 2	
Proportion Answering Correctly [95% CI]	.58 [.56, .59]
Difference versus Random Guessing (p-value)	.41 (<.001)
Mock Vignette Check 3	
Proportion Answering Correctly [95% CI]	.56 [.54, .58]
Difference versus Random Guessing (p-value)	.39 (<.001)

Notes: Flesch Reading Ease score obtained from https://datayze.com/readability-analyzer. Scores between 50 and 70 are approximately equivalent to high-school English; below 30 is college graduate-level English. MVC "Proportion Answering Correctly" reports logistic 95% confidence intervals (CI). "Difference versus Random Guessing (p-value)" based on results from a one-sample difference-in-proportions test, wherein the probability of randomly guessing the correct response equals 1/6 (.1667). Two-tailed p-value reported.

Mock Vignette 6 : "Mandatory Sentencing" (adapted from Gross (2008)).

Note: This was the MV & MVC used in MTurk 1 & Qualtrics Studies (see Section C)

Intro Screen
Next, we would like to ask you an additional question on a different topic. Please read the following news article.

Mock Vignette
Frederick Jackson, "The Case against Mandatory Minimum Sentencing"

It is now clear that mandatory minimums and their ripple effects are not punishing the major drug players they were intended for.

Instead we have a system where first time non-violent offenders can receive penalties greater than the average state sentence for murder or voluntary manslaughter. The federal mandatory minimums are determined by the amount of drugs—for example, a 10-year sentence is imposed for possession of 1,000 marijuana plants, while 5 grams of crack cocaine will send a defendant to

jail for five years. They fail to take account of whether the crime involved violence or whether there are mitigating circumstances.

Because those who are more involved have more information to trade, it is the drug users and those who are caught up with the actions of loved ones who are put into jail. Under mandatory minimums the prison population has exploded and prison costs are skyrocketing. The national crime rate has been dropping for seven years, yet more Americans are going to jail than ever before. The number of prisoners nationwide has more than tripled over the past 20 years, according to Justice Department statistics. More than half of these prisoners were locked up for non-violent crimes, most of them drug driven.

Mock Vignette Check [new screen; randomized response options; correct answer shaded]

According to the article you just read, mandatory minimum sentences for drug offenses do not take into account whether the crime involved:

- Money laundering
- Smuggling
- Violence
- Bribery
- Selling drugs to minors

Table A6. Mandatory Sentencing Mock Vignette and MVC Analytics

Mock Vignette	
Sample(s) Used	1. MTurk (Study 1) 2. Qualtrics
Word Count	212
Average Time Spent on Screen [95% CI]	**M-Turk:** 51.58 seconds [46, 57] **Qualtrics:** 64.96 seconds [60, 70]
"Flesch Reading Ease" Score	49.9
Mock Vignette Check	
Proportion Answering Correctly [95% CI]	**M-Turk:** .71 [.68, .75] **Qualtrics:** .64 [.60, .67]
Difference versus Random Guessing (p-value)	**M-Turk:** .51 (<.001) **Qualtrics:** .44 (<.001)

Notes: Flesch Reading Ease score obtained from https://datayze.com/readability-analyzer. Scores between 50 and 70 are approximately equivalent to high-school English; below 30 is college graduate-level English. MVC "Proportion Answering Correctly" reports logistic 95% confidence intervals (CI). "Difference versus Random Guessing (p-value)" based on results from a one-sample difference-in-proportions test, wherein the probability of randomly guessing the correct response equals 1/5 (.20). Two-tailed p-value reported.

TABLE A7. Summary Table of MVC Performance Across Experiments & Condition

	Percentage Answering # of MVCs Correctly			
	0	**1**	**2**	**3**
MTurk 1				
Control	28.43	71.57	--	--
Treatment	28.95	71.05	--	--
Overall	28.69	71.31	--	--
Qualtrics				
Control	34.72	65.28	--	--
Treatment	37.69	62.31	--	--
Overall	36.22	63.78	--	--
MTurk 2				
Control	11.97	15.96	35.41	36.66
Treatment	11.66	16.13	34.24	37.97
Overall	11.82	16.04	34.83	37.31
NORC				
Control	14.05	31.13	38.02	16.80
Treatment	11.75	28.46	38.90	20.89
Overall	12.87	29.76	38.47	18.90
Lucid				
Control	12.83	16.55	26.11	44.51
Treatment	12.65	15.25	25.25	46.84
Overall	12.74	15.91	25.69	45.67

Notes: Table displays % of each sample that passed a given number of mock vignette checks (MVCs). The MTurk1 and Qualtrics studies featured only 1 MVC, while all others featured 3 MVCs.

Table A8 below provides a series of protocols for constructing, implementing, and analyzing Mock Vignettes (MVs) and Mock Vignette Checks (MVCs). In addition to these protocols, there are several other considerations that may or may not be relevant to researchers. First, the majority of our studies implemented "forced" responses for MVCs (i.e., prevented "skipping over" an MVC without answering it. While this is the ideal practice given that skipping over questions forces the researcher to assume—rather than measure—inattentiveness, in some cases Institutional Review Boards (IRBs) may not allow the usage of "forced" responses. In such cases, the next best alternative would be instituting prompts for respondents to answer the MVC in the even they attempt to skip over it. If and when a respondent does skip over an MVC, it is reasonable to code that respondent as inattentive—i.e., as if they answered the MVC incorrectly (especially if a timer was used and indicates very little time spent on the question) as this will help preserve sample size. This is the strategy we employed in the NORC study.

Second, in designing the MVs, our general strategy was to first use Google News search to find local news stories about politics. We expect that local outlets would produce less sensationalist

TABLE A8. 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.

and more politically neutral content. From a list of recent stories covered by local outlets, we filtered out partisan topics (e.g., abortion, immigration) and instead selected topics such as stadium licensing that had no specific partisan valence in terms of issues or elected officials. We then used a readability analyzer (see table notes above) to ensure that the content was not overly sophisticated, and also made sure to avoid including any partisan, ideological, or emotional content

when paraphrasing the content produced in the news story. While we cannot be certain how such content might impact treatment receipt and/or treatment effect estimates, our principal aim was to ensure that the content remained neutral and benign in tone. A potential concern, for example, could be that the particular emotion aroused by an MV is correlated with one's outcome of interest, and/or may substantially impact how a treatment is received by respondents.

How similar should the MV be to the researcher's experiment? While we do not have the necessary data to answer this question, we believe that, to the extent that they can in the context of their own experiment, researchers may benefit from having the MV be roughly similar in length and general appearance to the experimental vignette (e.g., similar text size and font). Beyond that, however, we believe that intentionally making the MV similar to the experiment *in terms of content* (e.g., an MV about terrorism when the experiment is also about terrorism) runs the risk of inducing "pretreatment" effects (Druckman and Leeper 2012), which could undermine one's ability to detect significant differences between experimental groups.

APPENDIX B:
REPLICATED STUDIES & SAMPLE CHARACTERISTICS

This appendix contains the wording for the experimental treatments and outcome questions.
Note: Numeric coding values appear in parentheses beside response options.

Student Loan Forgiveness Experiment (MTurk 1, NORC, and Lucid Studies)

Control Condition
According to the U.S. Department of Education, college student loan debt now exceeds one trillion dollars, which surpasses the total credit card debt in the United States. This has led to proposals for a student loan forgiveness program.

Treatment Condition
According to the U.S. Department of Education, college student loan debt now exceeds one trillion dollars, which surpasses the total credit card debt in the United States. This has led to proposals for a student loan forgiveness program. A number of expert economic analysts suggest that a student loan forgiveness program would have serious negative effects on the economy. When individuals accept a student loan, they know they are required to pay it back. By transferring this individual responsibility and debt to the national government, the burden falls on all taxpayers and lets students avoid their financial obligations.

Outcome Measure
To what extent do you oppose or support the proposal to forgive student loan debt?

Strongly oppose (=1)
Oppose (=2)
Slightly oppose (=3)
Neutral / Neither oppose nor support (=4)
Slightly support (=5)
Support (=6)
Strongly support (=7)

KKK Demonstration Experiment (Qualtrics and Lucid Studies)

Free Speech Condition

"Ku Klux Klan (KKK) Plans to Demonstrate, Testing Commitment to Free Speech"
How far is Ohio State University (OSU) prepared to go to protect freedom of speech? The Ku Klux Klan has requested a permit to conduct a speech and rally on the OSU campus during the Fall of 2020. Officials and administrators will decide whether to approve or deny the request in September. Numerous courts have ruled that the U.S. Constitution ensures that the Klan has the right to speak and hold rallies on public grounds, and that individuals have the right to hear the Klan's message if they are

interested. Many of the Klan's appearances in the state have been marked by violent clashes between Klan supporters and counter demonstrators who show up to protest the Klan's racist activities. In one confrontation last October, several bystanders were injured by rocks thrown by Klan supporters and protesters. Usually, a large police force is needed to control the crowds.

Opinion about the speech and rally is mixed. Many students, faculty, and staff worry about the rally, but support the group's right to speak. Clifford Strong, a professor in the law school, remarked, "I hate the Klan, but they have the right to speak, and people have the right to hear them if they want to. We may have some concerns about the rally, but the right to speak and hear what you want takes precedence over our fears about what could happen."

Public Order Condition

"Ku Klux Klan (KKK) Plans to Demonstrate, Raising Public Safety Concerns"
Can campus police prevent a riot if the KKK comes to town? The Ku Klux Klan has requested a permit to conduct a speech and rally on the Ohio State University (OSU) campus during the Fall of 2020. Officials and administrators will decide whether to approve or deny the request in September.

Numerous courts have ruled that the U.S. Constitution ensures that the Klan has the right to speak and hold rallies on public grounds, and that individuals have the right to hear the Klan's message if they are interested. Many of the Klan's appearances in the state have been marked by violent clashes between Klan supporters and counterdemonstrators who show up to protest the Klan's racist activities. In one confrontation last October, several bystanders were injured by rocks thrown by Klan supporters and protesters. Usually, a large police force is needed to control the crowds.

Opinion about the speech and rally is mixed. Many students, faculty, and staff have expressed great concern about campus safety and security during a Klan rally. Clifford Strong, a professor in the law school, remarked, "Freedom of speech is important, but so is the safety of the OSU community and the security of our campus. Considering the violence at past KKK rallies, I don't think the University has an obligation to allow this to go on. Safety must be our top priority."

Outcome Measure

Would you support or oppose allowing the Ku Klux Klan (KKK) to demonstrate on the Ohio State University campus?

Strongly oppose (=1)
Oppose (=2)
Slightly oppose (=3)
Neutral / Neither support nor oppose (=4)
Slightly support (=5)
Support (=6)
Strongly support (=7)

Welfare Deservingness Experiment (MTurk 2 and Lucid Studies)

Unlucky Condition
People in the United States can have a variety experiences in life, including experiences that are related to one's ability to remain employed. Some people are able to remain continuously employed for long periods of time, perhaps even their entire lives; for other people, however, there are periods of time in which they are not employed. At present there is substantial discussion about federal policies related to unemployed persons.

Imagine a man who is currently on social welfare. He has always had a regular job, but has now been the victim of a work-related injury. He is very motivated to get back to work again.

Lazy Condition
People in the United States can have a variety experiences in life, including experiences that are related to one's ability to remain employed. Some people are able to remain continuously employed for long periods of time, perhaps even their entire lives; for other people, however, there are periods of time in which they are not employed. At present there is substantial discussion about federal policies related to unemployed persons.

Imagine a man who is currently on social welfare. He has never had a regular job, but he is fit and healthy. He is not motivated to get a job.

Outcome Measure
Regarding the man you just read about, to what extent do you disagree or agree that the eligibility requirements for social welfare should be *tightened for persons like him*?

Strongly disagree (1)
Disagree (2)
Slightly disagree (3)
Neutral / Neither agree nor disagree (4)
Slightly agree (5)
Agree (6)
Strongly agree (7)

Immigration Policy Experiment (Lucid Study)

Low-status Kuwaiti Individual Condition

Rashid Siddiqui is a native of Kuwait. He wants to come to the US and find a job as a construction worker. Eventually, he would like to settle in the US and become an American citizen. He is 30 years old and lives in Kuwait City. Rashid and his wife have two sons and one daughter. His father is in poor health and no longer able to work. Rashid helps pay for his parents' living expenses and also for the education of his two younger brothers and one sister. Rashid is a graduate of Khalifa School – a

vocational high school in Kuwait. After graduating, he has held various part-time jobs including construction worker, taxi driver, and house painter. He is learning English.

High-status Mexican Individual

Roberto Sanchez is a native of Mexico. He would like to come to the US to be an engineer. Eventually, he would like to settle in the US and become an American citizen. He is 30 years old and lives in Mexico City. Roberto and his wife have two sons and one daughter. His father is in poor health and no longer able to work. Roberto helps pay for his parents' living expenses and also for the education of his two younger brothers and one sister. Roberto received his undergraduate degree in structural engineering at Universidad Tecnológica de México. After graduating, he was hired by Polywell Computers and has worked at Polywell Computers as a quality assurance technician. He is learning English.

Outcome Measures (combined into an additive scale)

The individual you just read about is applying for a permit to work in the U.S. Given what you know about him, do you think his application for a work permit should be approved or rejected?

Approved (=2)
Rejected (=0)
Cannot Say (=1)

If his application were approved, for how long should he be permitted to work?

6 months (=0)
1 year (=1)
2 years (=2)
3 years (=3)

Assume that the individual you read about comes to the U.S. on a work permit and then he decides to apply for American citizenship. Do you think his citizenship application should be approved or rejected?

Approved (=2)
Rejected (=0)
Cannot say (=1)

TABLE B1. Sample Characteristics of All Studies

	MTurk 1 (N=603)	Qualtrics (N=1040)	NORC (N=744)	MTurk 2 (N=804)	Lucid 1 (N=6,216)	Lucid 2 (N=9,148)
Median Income	50k-75k	20k-50k	50k-60k	50k-75k	25k-50k	40k-45k
Median Education.	College Grad	Some College	Some College	College Grad	Some College	Some College
Mean Age	36.82	46.43	48.84%	38.38	48.28	48.41
Female	51.74%	49.52%	51.02%	50.62%	44.31%	51.22
White	71.48%	62.31%	67.51%	74.38%	74.63%	74.08%
Black	7.63%	11.83%	10.34%	8.58%	10.50%	11.88%
Hispanic	11.77%	17.31%	14.93%	8.33%	7.12%	15.58%
Democrat	52.07%	46.54%	46.33%	52.86%	43.63%	44.39%
Independent	15.26%	19.71%	17.34%	17.04%	18.88%	20.40%
Republican	32.67%	33.75%	36.34%	30.10%	37.49%	35.21%
Liberal	49.92%	35.29%	-	47.76%	32.81%	32.37%
Moderate	15.75%	29.81%	-	21.14%	32.37%	33.29%
Conservative	34.33%	34.90%	-	31.09%	34.82%	34.35%
Mean Political Interest	3.50	3.46	-	3.44	3.38	3.26

Notes: MTurk 1 study fielded via Amazon.com's Mechanical Turk in May 2019. NORC study fielded via NORC at the University of Chicago's Amerispeak Omnibus survey in November of 2019; Qualtrics study fielded via Qualtrics in August of 2019; MTurk 2 study fielded via Amazon.com's Mechanical Turk in January of 2020; Lucid 1 study fielded via Lucid in February of 2020. Lucid 2 study fielded in August of 2021. All studies fielded online. NORC study is a national probability sample of adults (additional information can be found here: https://amerispeak.norc.org/about-amerispeak/Pages/Panel-Design.aspx). Sampling for the Qualtrics and Lucid included quotas to mirror U.S. Census data on Age (18-24; 25-34; 35-44; 45-54; 55-64; 65+), Race/Ethnicity (Non-Hispanic White; Non-Hispanic Black; Hispanic; Asian; Other), and Geographic Region (West; Midwest; Northeast; South). Partisan groups (i.e., Democrats and Republicans) include those who report "leaning" toward one party. Political Interest measured on a five-point scale ranging from "Not interested at all" (1) to "Extremely interested" (5).

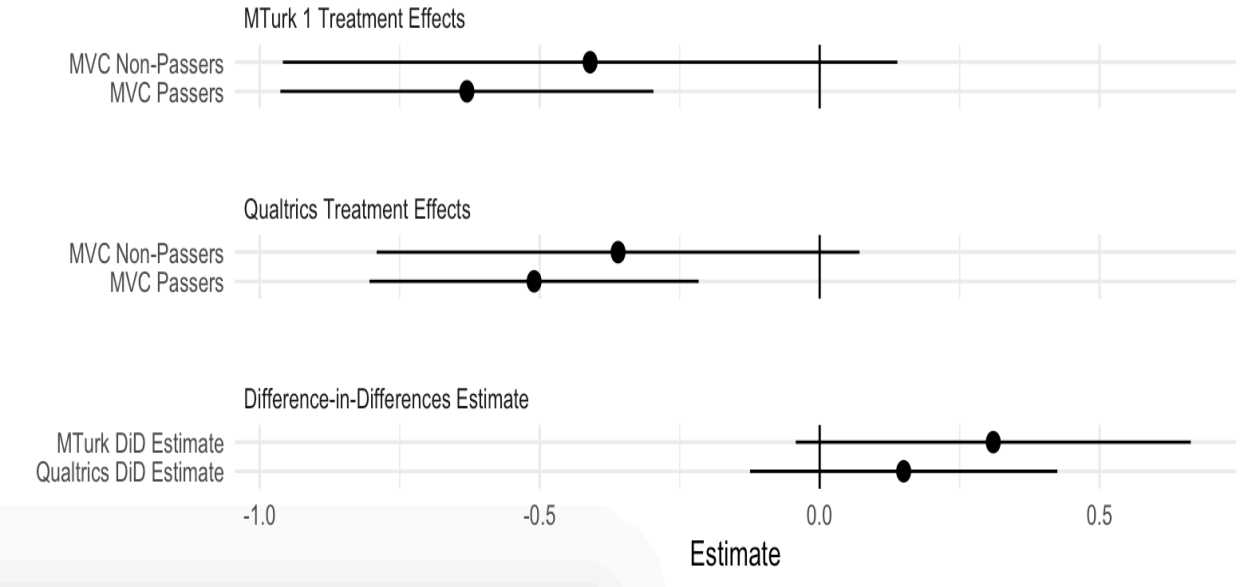
APPENDIX C:
RESULTS FOR MTURK 1 & QUALTRICS STUDIES

For both the MTurk 1 study and Qualtrics study, Figure C1 plots: (1) the treatment effect among MVC non-passers, (2) the treatment effect among MVC passers, and (at the bottom) (3) the absolute difference between these two estimates.

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, is significant at the p<.10 level (two-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 passers (-.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; p<.01). 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, though correctly signed, was not statistically significant (p=.31). Nevertheless, with the treatment effect for the sample as a whole being equal to -.46, this study again illustrates how neglecting inattentiveness will tend to yield weaker treatment effect estimates. This latter estimand is therefore akin to the average effect of receipt for compliers (AERC [see Harden, Sokhey, and Runge 2019 Supplemental Appendix pp.10-11]).

FIGURE C1. TREATMENT EFFECTS & DID ESTIMATES (MTURK 1 & QUALTRICS)



Notes: Figure displays treatment effect estimates stratified by MVC performance in MTurk 1 & Qualtrics studies, as well as the difference-in-differences estimate for both studies. 95% confidence intervals shown.

APPENDIX D:
DEMOGRAPHIC PREDICTORS OF MVC PERFORMANCE

We explored correlates of MVC performance across our five studies (see Table G1). Overall, the only consistent predictors of performance were (1) race (in particular, African-American and Hispanic identification), and (2) age. Nonwhite respondents tended to perform slightly less well than White respondents. Older respondents tended to perform somewhat better than younger respondents. However, these variables obtained only weak pairwise correlations with MVC performance. Racial variables correlate with MVC performance at < |.2|; age correlates with MVC performance at <|.33|. Further, when we ran the NORC, MTurk 2, and Lucid 1 studies (which featured the scaled MVC measure) with controlled interactions for all significant predictors of attentiveness (i.e., each significant demographic predictor X treatment), point estimates of the *treatment* X *MVC performance* interaction term did not substantively change, nor did the *p*-values for those interaction terms.[1] Lastly, we examined the degree to which the sample composition of the MTurk and Qualtrics studies changed with respect to race and age once we subsetted on MVC passers. For race, these changes amounted to a few percentage-points or less, while for age, the compositional changes amounted to approximately 3 years (49 versus 46 for the (Qualtrics) sample as a whole) and 1 year (38 versus 37 for the (MTurk 1) sample as a whole).

Gender obtains conventional significance ($p < .05$) in two of the five studies, with females performing slightly better than males. Education also obtains conventional significance in two studies, but its sign is inconsistent across the five studies. Notably, political variables (party identification, ideological self-placement, and political interest) were rarely significant predictors, and were inconsistently signed across the five studies. Taken together, these results suggest only minor changes in demographic composition when analyzing the attentive and, perhaps more importantly, little consequence for the CATE estimates.

As noted in the manuscript, because attentiveness is unlikely to be randomly distributed in the population, analyzing attentive respondents stands to alter the demographic composition of the sample. To reiterate, because the MVCs appear *pre*-treatment, any change in demographic composition will not, in expectation, yield biased estimates of treatment effects. Rather, it may simply limit the generalizability of one's findings to a broader population (cf. Coppock, Leeper, and Mullinix 2018; Mullinix et al. 2015).

[1] We conducted the same procedure for the MTurk 1 and Qualtrics studies discussed above; that is, we specified an interaction between treatment and MVC performance along with controlled interactions between treatment and other significant ($p<=.10$) predictors of MVC performance, and compared the treatment X MVC performance estimate to this estimate when no controlled interactions were included in the model. The MTurk 1 study saw little change in coefficient. The Qualtrics study saw a change from -.15 to -.10. However, given the relatively small samples and use of only one MVC rather than a scale, SEs were quite large relative to the point estimates in these models.

TABLE D1. Demographic Predictors of MVC Performance

	Mock Vignette Check (Binary)		Mock Vignette Check (0-1 Scale)		
	MTurk 1	**Qualtrics**	**NORC**	**MTurk 2**	**Lucid**
Female	0.06	0.06†	0.03	0.07**	0.04***
	(0.04)	(0.04)	(0.02)	(0.02)	(0.01)
African-American	-0.18**	-0.13*	-0.14***	-0.10*	-0.09***
	(0.07)	(0.06)	(0.04)	(0.04)	(0.01)
Hispanic	-0.26***	-0.02	-0.12***	-0.16***	-0.06***
	(0.06)	(0.05)	(0.03)	(0.04)	(0.02)
Asian	-0.09	-0.13†	0.05	-0.03	-0.06**
	(0.08)	(0.08)	(0.07)	(0.05)	(0.02)
Other	-0.19*	-0.02	-0.01	-0.24***	-0.03†
	(0.09)	(0.10)	(0.06)	(0.07)	(0.02)
Age	0.38***	0.48***	0.02	0.32***	0.48***
	(0.11)	(0.09)	(0.05)	(0.06)	(0.02)
Income	0.13	-0.04	0.13*	0.07	-0.05**
	(0.09)	(0.08)	(0.05)	(0.05)	(0.02)
Education	-0.11	0.11	0.27***	-0.12†	0.08***
	(0.11)	(0.09)	(0.08)	(0.07)	(0.02)
Political Interest	0.02	0.08	--	-0.06	0.05***
	(0.07)	(0.06)		(0.05)	(0.02)
Party ID	-0.11	-0.02	-0.01	-0.02	0.05**
	(0.09)	(0.06)	(0.04)	(0.05)	(0.02)
Ideology	-0.15†	-0.09	--	-0.08	0.02
	(0.09)	(0.07)		(0.05)	(0.02)
Constant	0.75***	0.43***	0.31***	0.66***	0.42***
	(0.08)	(0.07)	(0.06)	(0.05)	(0.01)
N	603	784	742	804	11,056
R-squared	0.11	0.07	0.09	0.11	.12

Notes: The table reports regression coefficients with standard errors in parentheses. To ease interpretation of results across the four studies, all models are OLS and the "Scale" outcome measures are recoded to range from 0 to 1. Political Interest ranges from 1=Not at all interested to 5=Extremely interested. All gender and racial identification variables are dichotomous; all continuous variables are recoded to range from 0 to 1. "Party ID" and "Ideology" are coded as follows: 1=Strong D/Extremely Liberal; 2=D/Liberal; 3=Lean D/Slightly Liberal; 4=Independent/Moderate; 5=Lean R/Slightly Conservative; 6=R/Conservative; 7=Strong R/Extremely Conservative. Income is coded as follows in all studies except NORC: 1=$0-$25k, increasing in $25k increments to 5; 5=$100k-$150k; 6=$150-$200k; 7=Over $200k. NORC: 1=<$5k increasing in $5k increments to 9, and then $10k increments to 11, then $15k increments to 14, then $25k increments to 18; 18=$200k or more. Education is measured in MTurk, Qualtrics and Lucid studies as: 1=Less than HS; 2=HS graduate; 3=Some college; 4=College degree; 5= Master's

degree; 6=Higher degree. Education in the NORC study is measured as 1=No formal education; 2=1st to 4th grade; 3=5th or 6th grade; 4=7th or 8th grade; 5/6/7/8=9th/10th/11th/12th grade. 9=HS diploma; 10=Some college; 11=Associate's degree; 12=Bachelor's degree; 13=Master's degree; 14=Professional or doctorate degree. The NORC study did not include measures of "Political Interest" or "Ideology." The Lucid model includes mock vignette and round fixed effects, and standard errors are clustered by respondent. *** $p<0.001$, ** $p<0.01$, * $p<0.05$, † $p<0.10$ (two-tailed).

APPENDIX E: VALIDATING MVCS USING TIMERS & FMCS

This section features results of our investigation into the validity of MVCs as a measure of attentiveness. We analyze the relationship between MVC performance and (1) screen timers (on the MV, experimental vignettes, experimental outcome, and total survey duration), and (2) factual manipulation checks (FMCs), which appeared after the experimental outcome measure(s). In the case of timers, we log-transform each timer, and then regress it onto either a binary (MTurk 1 and Qualtrics) or continuous-scale (NORC, MTurk 2, and Lucid) measure of MVC performance (recoded to range from 0 to 1). The resulting estimate therefore indicates the % change in time spent given an increase from 0 to 1 in MVC performance. In the case of FMCs, we code responses to these items as either incorrect (0) or correct (1), generating a variable indicating performance on the FMC. We then specify a logistic regression model, which regresses the binary FMC performance measure onto MVC performance. This approach enables us to obtain the change in Pr(FMC=Correct) given a one-unit increase in MVC performance.

TABLE E1. Mock Vignette Check (MVC) Passage Predicts Greater Attentiveness to Experiment

	%Δ Time Spent					Δ Probability
	Mock Vignette	*Vignette (Control)*	*Vignette (Treatment)*	*Outcome Measure*	*Survey Duration*	*Pass FMC*
Maximal Effect of MVC						
Student Loan Experiment (MTurk 1)	123%*	63%*	108%*	14%*	36%*	.49*
KKK Experiment (Qualtrics)	119%*	132%*	110%*	28%*	41%*	.35*
Student Loan Experiment (NORC)	164%*	68%*	122%*	-14%	88%*	.35*
Welfare Experiment (MTurk 2)	196%*	141%*	204%*	79%*	57%*	.45*

Notes: The key independent variable is mock vignette check (MVC) performance, which is binary for the MTurk 1 and Qualtrics studies, and continuous (recoded to range from 0 to 1) for the NORC and MTurk 2 studies. Each column represents a different outcome measure of interest. Figures for "%Δ Time Spent" outcomes represent % change in time spent on a given outcome, and were generated from log-linear OLS regression models wherein the amount of time was log-transformed. Figures for "Pass FMC" outcome represent change in probability of correctly answering the factual manipulation check (FMC), and were generated from a logistic regression model. "[Control/Treatment] Vignette" = the Free Speech/Public Order frame in the *KKK* experiment, and the Unlucky/Lazy frame in the *Welfare* experiment. * significance at p<.05 or lower (one-tailed).

TABLE E2. Mock Vignette Check (MVC) Passage Predicts Greater Attentiveness to Experiment (Lucid 1 Study)

	%Δ Time Spent					Δ Probability
	Mock Vignette	*Vignette (Control)*	*Vignette (Treatment)*	*Outcome Measure*	*Survey Duration*	*Pass FMC*
Maximal Effect of MVC						
Student Loan Experiment (n=1264)	241%*	101%*	178%*	51%*	69%*	.41*
KKK Experiment (n=1264)	206%*	201%*	186%*	51%*	70%*	.59*
Welfare Experiment (n=1243)	282%*	177%*	174%*	72%*	71%*	.57*
Immigration Experiment (n=1267)	261%*	198%*	189%*	69%*	74%*	.68*

Notes: The key independent variable is mock vignette check (MVC) performance, which is a continuous measure indicating moving from answering 0 MVCs correctly to answering all 3 MVCs correctly for the second-round MV. Each column represents a different outcome measure of interest from the second round. Figures for "%Δ Time Spent" outcomes represent % change in time spent on a given outcome, and were generated from log-linear OLS regression models wherein the amount of time was log-transformed. Figures for "Pass FMC" outcome represent change in probability of correctly answering the factual manipulation check (FMC), and were generated from a logistic regression model. "Treatment Vignette" = the Public Order frame in the *KKK* experiment, the Lazy frame in the *Welfare* experiment, and High-Status Mexican frame in the *Immigration* experiment. For "outcome measure" in *Immigration* experiment, all three outcome measure timers were combined. All models control for which MV was observed in second round. * significance at $p<.05$ or lower (one-tailed)

APPENDIX F:
MVC PLACEMENT, CATE SIZE & EFFECTS ON ATTENTIVENESS

Here we first display the CATEs from models underlying Figure 3 in the manuscript.

TABLE F1. Results Underlying Figure 3 (Lucid Data)

	Student Loan Forgiveness	KKK Demonstration	Welfare Deservingness	Immigration Policy
Treatment	0.18*	0.20**	0.52***	0.19**
	(0.07)	(0.07)	(0.07)	(0.07)
MVC Performance	0.04^	0.02	-0.24***	0.08**
	(0.03)	(0.03)	(0.02)	(0.02)
Treatment X MVC Performance	0.10**	0.12***	0.36***	0.08**
	(0.03)	(0.03)	(0.03)	(0.03)
Scientific Publishing	-0.25**	-0.09	0.15^	-0.12
	(0.09)	(0.09)	(0.08)	(0.08)
Stadium Licenses	-0.19*	-0.13	0.17*	-0.17*
	(0.09)	(0.09)	(0.08)	(0.08)
Sulfur Reductions	-0.24**	-0.22*	0.20*	-0.15^
	(0.08)	(0.09)	(0.08)	(0.08)
Hazardous Plants	-0.18*	-0.15^	0.14^	-0.07
	(0.09)	(0.09)	(0.08)	(0.08)
Round 2	0.12**	-0.04	0.03	0.01
	(0.04)	(0.04)	(0.04)	(0.04)
Constant	0.05	0.11	0.27***	-0.03
	(0.07)	(0.07)	(0.07)	(0.07)
N	2,755	2,729	2,742	2,743
R-squared	0.05	0.05	0.30	0.05
Adjusted R-squared	0.0496	0.0474	0.300	0.0473

Notes: Lucid data. Coefficients are OLS, with SEs clustered by respondent. *** p<.001; **p<.01; *p<.05; ^ p<.10 (two-tailed).

We next investigate whether CATEs significantly depend on the MVs that are assigned. In other words, we evaluate whether certain MVs outperform others with respect to being able to recover larger CATEs. Using the Hazardous Plants MV as our baseline, Table F1 indicates the interaction between treatment status and MVC performance is statistically significant (p<.001). The following three entries display triple difference estimates that allow us to test whether the effect size of the interaction term varies based on the assigned MV. These triple difference estimates range from differences of half a percentage point to two percentage points. Thus, differences between MVs—in terms of predicting larger CATEs—to be minimal and not statistically discernible from zero.

TABLE F2. Heterogeneity in Conditional Average Treatment Effects by Mock Vignette

	Experimental Outcome Measure
Treatment × MVC Score	.222*
	(.038)
Treatment × MVC Score × Scientific Publishing MV	-.020
	(.056)
Treatment × MVC Score × Stadium Licenses MV	-.008
	(.054)
Treatment × MVC Score × Sulfur Dioxide MV	-.022
	(.055)
N	10,969

Notes: Lucid 1 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. Constituent terms suppressed to simplify presentation of triple difference estimates. *p<0.05 or lower (one-tailed)

By virtue of its design, the Lucid 1 study also enables us to also examine consequences of an MV's placement 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 examine whether the *treatment* X *MVC* interaction (i.e., the CATE) changes in magnitude as a result of the MV appearing directly (versus long) before the (second-round) experiment. Specifically, we assess whether the placement of MVs predicts *larger* CATEs, under the assumption that responses to MVs placed directly before the experiment should be more indicative of attention in that moment of the experiment than responses to MVs placed long before the experiment.

As per Table F2, though CATEs are larger when comparing Round 2 to Round 1 MVs, this difference corresponds to .7% of a control group standard deviation. We formally test whether these differences are statistically discernible from zero using an F-test, and fail to reject the null of parameter equality ($F(1, 4280) = .01$, $p = .91$).

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.[2]

[2] 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

TABLE F3. Heterogeneity in Conditional Average Treatment Effects by Mock Vignette

	Experimental Outcome Measure
Treatment (Round 2)	.033
	(.075)
MV (Round 1)	-.071
	(.025)
MV (Round 2)	-.020
	(.026)
Treatment (Round 2) × MV (Round 2)	.132*
	(.035)
Treatment (Round 2) × MV (Round 2)	.139*
	(.038)
N	4,826

Notes: Lucid 1 study. OLS regression coefficients. Outcome is standardized within each experiment (control group standard deviations). Mock Vignette Check Score ranges from 0 to 3. *p<0.05 or lower (one-tailed)

Finally, we do find some evidence that using (versus not using) an MV is associated with modestly better performance on correctly answering FMCs, suggesting that MV/MVCs may also encourage slightly greater attentiveness to one's experiment. Specifically, using a logistic regression model that controls for experiment (with SEs clustered by respondent), we find that featuring an MV increases the probability of correctly answering the first-round FMC by 5.8 percentage points (p<.001) in the Lucid 1 study, and by 2.3 percentage points (p=.05) in the second Lucid study.

performance in round 1, suggesting that attentiveness levels shortly before (versus longer before) the experiment more closely resemble attentiveness during the experiment.

APPENDIX G:
TESTING WHETHER MOCK VIGNETTES DISTORT TREATMENT EFFECTS

The results of this investigation appear in Figure G1. 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 did and did not observe the MV. In the NORC sample, 27% of the sample was not shown an MV (*n*=279). In this study, there was also no statistically distinguishable difference in treatment effect estimates between the no MV and MV condition. In the Lucid 1 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. Lucid 2 also randomly varied inclusion of the MV before subjects were randomly assigned to two of four experiments (*n=1,063)*. This provides us with four additional opportunities to assess if exposure to a MV augments or decreases effect sizes.

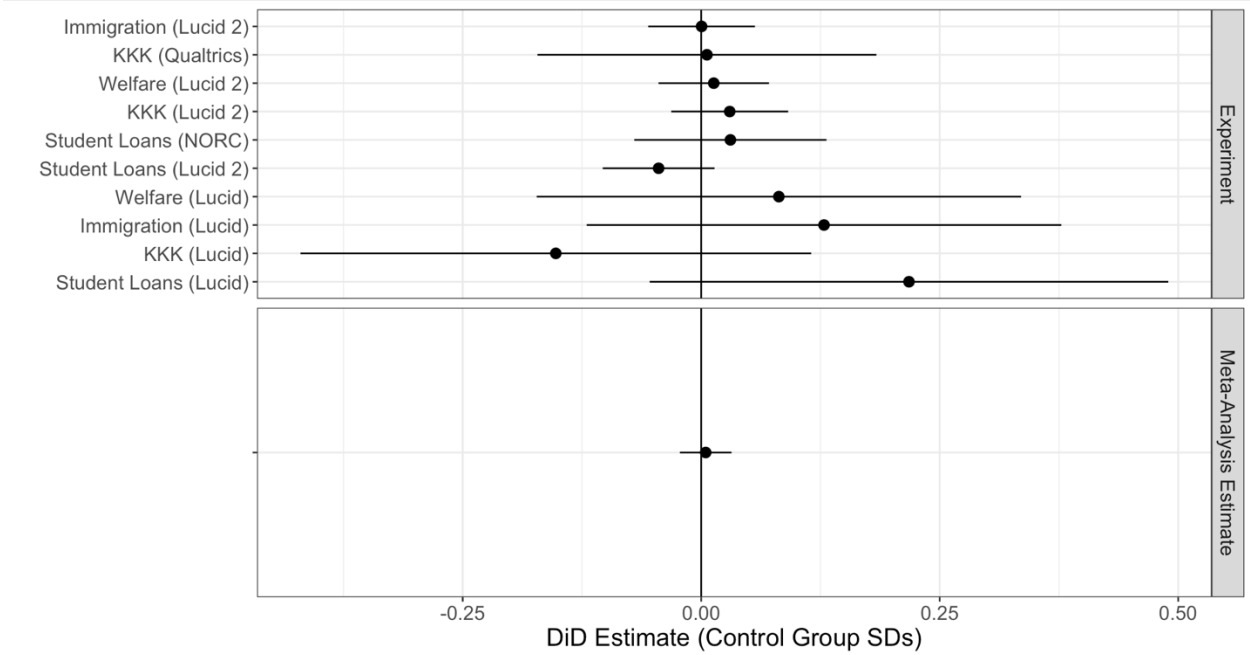
As demonstrated in Figure 4, we find no evidence that those respondents who observed, versus did not observe, an MV before the experiment exhibited significantly different treatment effects. Treatment effects were, in each experiment, substantively and statistically similar across these two groups. Indeed, the DiD estimates are statistically indistinguishable from zero in all four cases.[3]

Moreover, the sign on the DiD estimates is inconsistent—that is, in two instances the sign is opposite the ITT estimate (the KKK and Student Loans study), but in the other 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 find that the "MV inclusion effect" is negligible (.005 control group standard deviations), precisely estimated (SE = .01), and also not statistically distinguishable from zero.[4]

[3] Each of these four analyses had between 1,239 and 1,270 respondents in total, making it unlikely that such results are simply due to insufficient power.

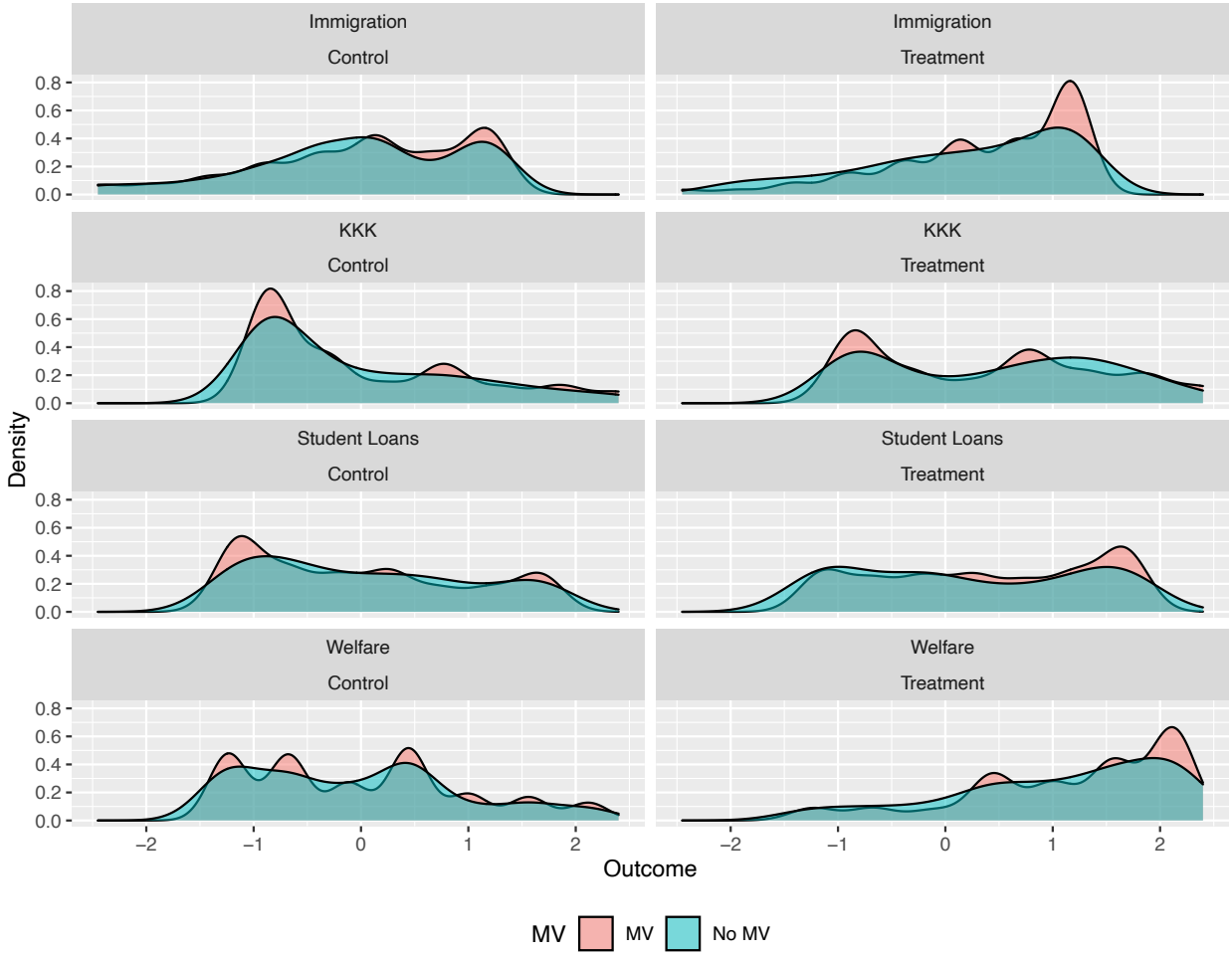
[4] In contrast to fixed-effect meta-analysis, which assumes that studies are estimating a single "true" effect, random-effects meta-analysis models assume that effects are drawn from a larger population, and may vary from study to study. In our case, the fixed-effect meta-analysis estimate (.0047) is identical to the random-effect estimate (.0047).

FIGURE G1: No Significant Change in Treatment Effects When a Mock Vignette Is Used



Secondly, we used each experiment in Lucid data to study whether usage of an MV may result in significantly different *variances* in the outcome measure. In other words, while MV usage does not (as shown above) distort the treatment effect *on average*, it is possible that it could lead to heterogeneous effects (e.g., increasing treatment effects for some individuals, and decreasing effects for others, versus if no MV had been used). To the extent this is the case, we should see that the variances in the outcome measure are significantly different depending upon whether an MV was seen. We tested the possibility of different outcome variances not only within each experiment, but also within each experimental condition. The results are shown in Figure G2, and we formally tested for significant differences in variances within each panel of Figure G2. In no instance did we find that the group seeing an MV had a significantly different variance than the group that did not see an MV ($p > .10$ in all cases).

FIGURE G2: No Significant Change in Outcome Variance Due to Mock Vignette Use



APPENDIX H:
SUBSETTING ON MVC PERFORMANCE & DETECTING SIGNIFICANT EFFECTS

MVCs enable a researcher to analyze experimental treatment effects among a more (versus less) attentive sub-sample of respondents. However, by reducing the size of the sample one is analyzing, statistical uncertainty increases (*ceteris paribus*) and, thus, the statistical power needed to detect a statistically significant effect (e.g., at the conventional p<.05 level) potentially decreases. On the other hand, as a larger treatment effect is likely to be detected among the attentive subsample, it is not necessarily the case that statistical power will decline, nor, more broadly, that one will obtain a non-significant treatment effect when analyzing the attentive.

To investigate these concerns more directly, we analyze how the t-statistic changes in each of our replicated experiments as we analyze an increasingly attentive sub-sample (i.e., as we examine better MVC performance). Specifically, we regress each study's outcome onto each experiment's binary treatment indicator, yielding an OLS coefficient (the TE), and then record how the *t*-statistic on this coefficient changes as we move from >=0 MVCs correct, to >=1 MVCs correct, to (if applicable) >=2 MVCs correct, to (if applicable) =3 MVCs correct (using the MV and MVC featured in that particular study).

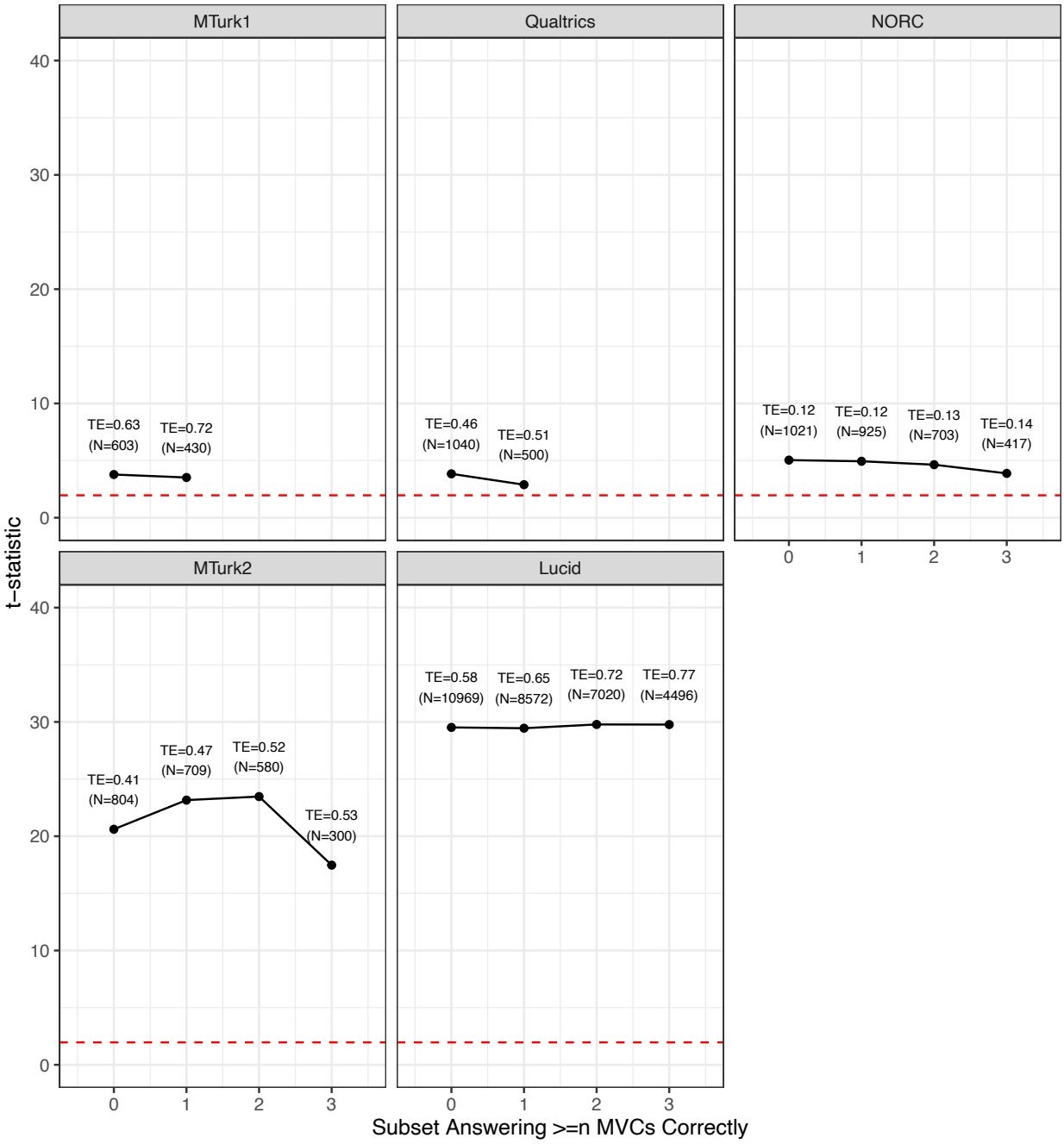
As *t* is a function of the estimated treatment effect (TE) divided by the standard error (SE), change in this statistic is a useful indicator of the net consequence (on our ability to detect a statistically significant effect) of 1) a larger effect, yet 2) smaller sub-sample.

Figure H1 displays the results of these analyses (absolute values of *t* and TEs shown to simplify presentation). The dashed red horizontal line indicates a *t*-statistic of 1.96, which, for large samples, yields a (two-tailed) *p*-value of .05. The *t*-statistic for the ITT is indicated by >=0 MVCs correct (i.e., the first point on each *x*-axis). TEs and *n* sizes are shown for each value of *t*.

There are several noteworthy features of these figures. First, it always the case that we observe a larger TE as we move to the right along the x-axis—i.e., as we analyze a more attentive sub-sample of respondents, regardless of experiment. Second, it is not always the case that we observe a substantial decrease in *t* as we analyze a more attentive sub-sample: the MTurk 2 and Lucid studies show *increases* in *t*, while the NORC study shows negligible declines in *t*. Third, and perhaps most importantly, in no case do we observe that a statistically significant ITT (i.e., at >=0 MVCs correct) becoming non-significant (i.e., *p*>.05, two-tailed) among a more attentive sub-sample. That is, TEs remain statistically significant even among the most attentive sub-sample of respondents, despite this group being substantially smaller than the sample as a whole.

In sum, in each of the experiments we replicated, we do not find it to be the case that analyzing a more attentive sub-sample (as measured by MVC performance) will yield a non-significant (albeit larger) treatment effect. Again, this is partly due to the fact that we consistently find a larger treatment effect among the more attentive, which helps to offset the loss of power due to smaller sample size. And while this offset may not always be large enough to yield a *larger t*-statistic for the attentive, our results indicate that the researcher can nevertheless uncover a statistically significant treatment effect even among the most attentive sub-sample of respondents.

FIGURE H1: Changes in *t* with Better MVC Performance



Notes: Absolute values of *t*-statistics and TEs shown. Sample and sub-sample sizes shown in parentheses. Red horizontal line indicates $t = 1.96$. The TE at 0 on the *x*-axis is equivalent to the ITT for the sample as a whole. MTurk 1 and Qualtrics studies only featured one MVC; all others featured 3 MVCs.

APPENDIX I:
COMPARING MVCS AND INSTRUCTIONAL MANIPULATION CHECKS

In August of 2021, we fielded a separate, pre-registered study fielded via Lucid (total n=9,000) that randomly assigned respondents to answer either three IMCs or three MVCs at the start of the survey (whether the three IMCs or three MVCs appeared first was determined randomly, and both sets of these items were combined into additive scales for analysis).[5] Pre-registration details can be found at https://osf.io/2zp5m/?view_only=ec3ae68098964d27bcdaf1aeb31edf5a. Respondents were then randomly assigned to two of the four experiments featured in the main manuscript (all vignettes, outcome measures, and factual manipulation checks (FMCs) remained the same as in the previous Lucid experiment). This design enables us to compare mock vignette checks (MVCs) to instructional manipulation checks (IMCs) using differences in conditional average treatment effects (CATEs), response timers duration measures (RTs), and post-outcome factual manipulation check (FMC) performance. We also compare the demographic profiles of IMC and MVC passers.

TABLE I1. MVC Performance, IMC Performance, and CATE Estimates (Overall)

	MVC Model	IMC Model
Treatment	0.088***	0.192***
	(0.029)	(0.029)
MVC Score	-0.298***	
	(0.028)	
Treatment x MVC Score	0.690***	
	(0.041)	
IMC Score		-0.164***
		(0.029)
Treatment x IMC Score		0.530***
		(0.042)
Intercept	0.173***	0.093***
	(0.020)	(0.019)
Observations	16,156	16,156

Notes: Dependent variable is in control group SD units. *p<0.1; **p<0.05; ***p<0.01

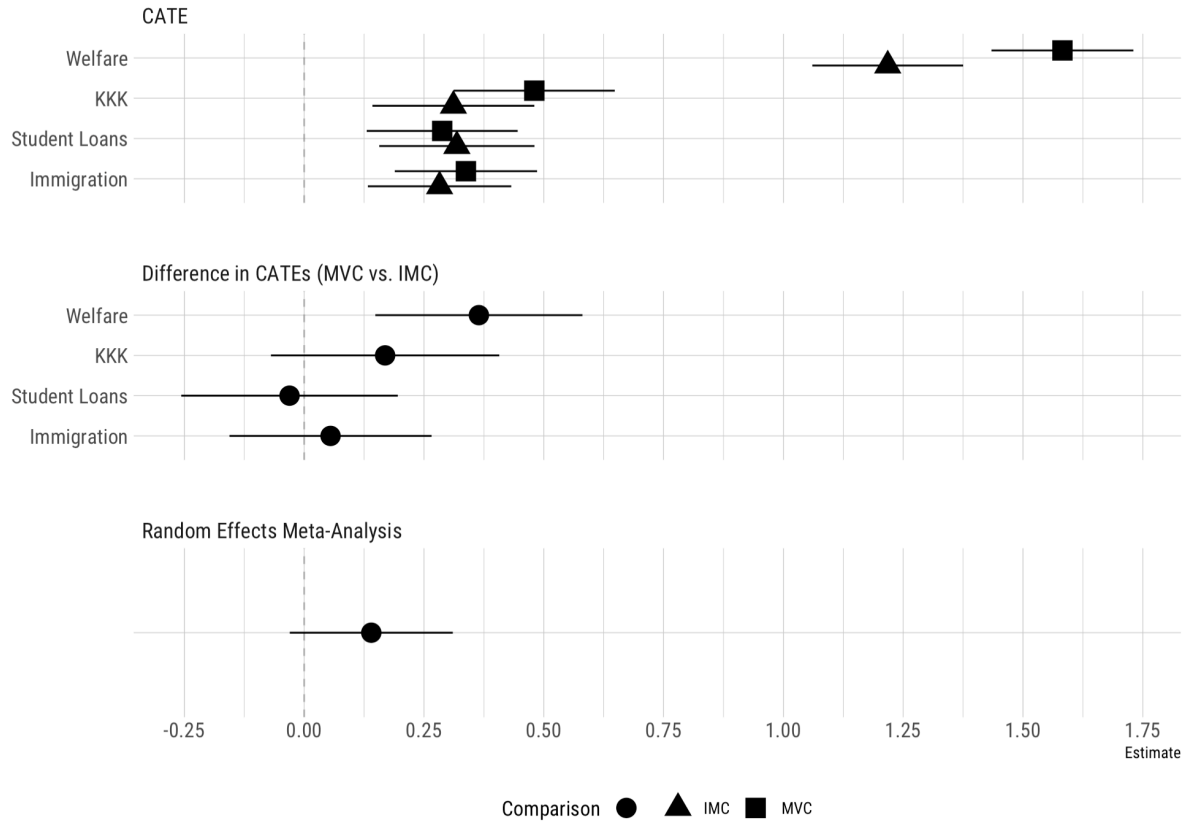
We begin with comparing MVC and IMC CATEs. Table I1 shows that pooling across experiments, the difference in treatment effects between the least and most attentive, as measured by the MVC, is .69 control group standard deviations (SE = .04; p < .05). For the IMC measure, the difference in treatment effects between the least and most attentive is .53 control group standard

[5] The specific IMCs that were used were (1) preferred news website, (2) how respondent is currently feeling, and (3) favorite color (see Berinsky, Margolis and Sances 2014, including the authors' online supplemental material).

deviations (SE = .04; $p < .05$). Both measures significantly predict treatment effect heterogeneity—i.e., better performance on the MVC and IMC scales is associated with larger treatment effects. However, MVCs slightly outperform IMCs, with a difference of .16 standard deviation units in CATEs (SE = .059; $p < .05$).

Because pooling across experiments assumes a single effect size, we also present a comparison of CATEs *within* each experiment. As shown in Figure I1, the MVC CATE is larger than the IMC CATE in 3 out of 4 cases, and the difference between the two measures is statistically significant in the welfare experiment (Δ =.365; SE = .110; $p < .05$). The meta-analytical estimate of the difference between MVC and IMC CATEs is .14 (SE = .087; p = .11). Therefore, we find that MVCs and IMCs perform similarly in conditioning treatment effects, though MVCs display a slight, but fairly consistent advantage over IMCs.

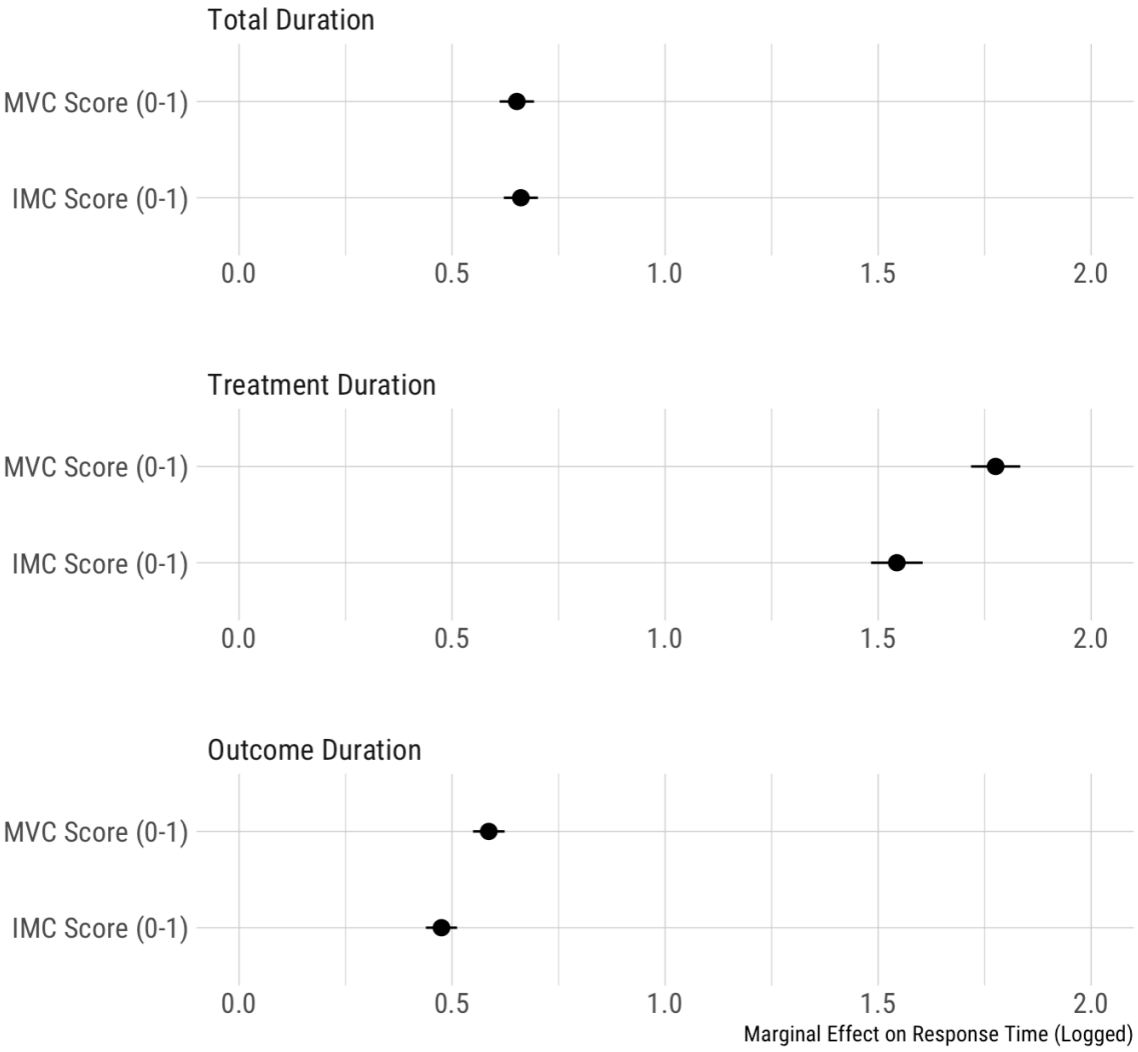
FIGURE I1. MVC Performance, IMC Performance, and CATE Estimates (By Experiment)



Notes: 95% CIs shown. Dependent variable is in control group standard deviation units.

We now turn to a comparison of MVCs and IMCs with respect to response times and study duration. To do this, we log each duration measure and estimate separate bivariate OLS regressions of duration on MVC and IMC scores. Per Figure I2, we find that MVCs and IMCs perform nearly

FIGURE I2. MVC Performance, IMC Performance, and Response Time Measures

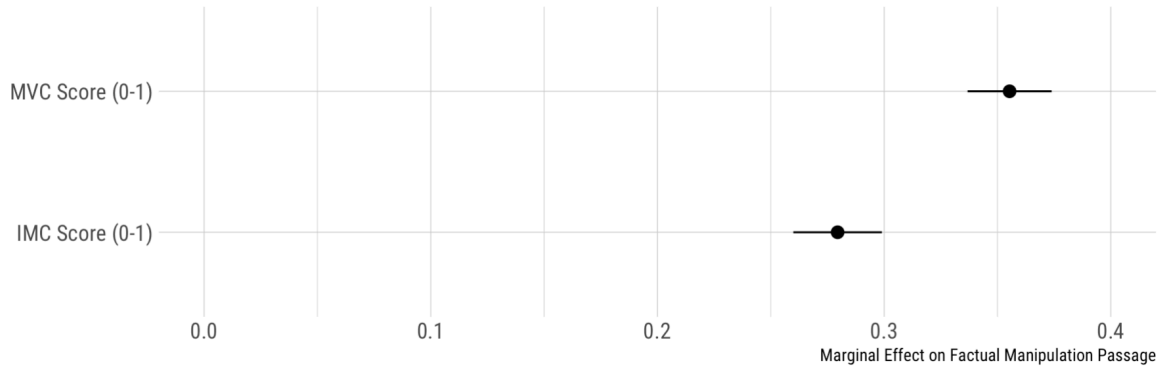


Notes: 95% CIs shown. All outcome measures are log-transformed.

identically in predicting total study duration (β_{mvc} = .65; β_{imc} = .66).[6] However, MVCs predict substantially more time spent on experimental stimuli and outcome measures. The marginal effect of MVC scores on logged experimental stimuli duration is 1.78 (SE = .03), which translates to a 178% increase moving from 0 correct MVCs to 3 correct MVCs. By comparison, the estimated marginal effect of IMC scores was smaller in size at 1.54 (SE = .03). This difference is statistically significant (Δ =.24; SE = .04; p < .05). The marginal effect of MVC scores on logged outcome measure duration is .59 (SE = .02), compared to .48 (SE = .02) for IMC scores. This difference is also statistically significant (Δ =.11; SE = .03; p < .05). In sum, we find evidence that MVCs outperform IMCs in predicting more time on experimental stimuli and outcome variables, though not total study duration.

[6] When we subtract the time spent on MVCs and IMCs from each respondent's total survey duration, we find extremely similar results, though with MVCs displaying a slight advantage over IMCs.

FIGURE I3. MVC Performance, IMC Performance, and FMC Passage Rates

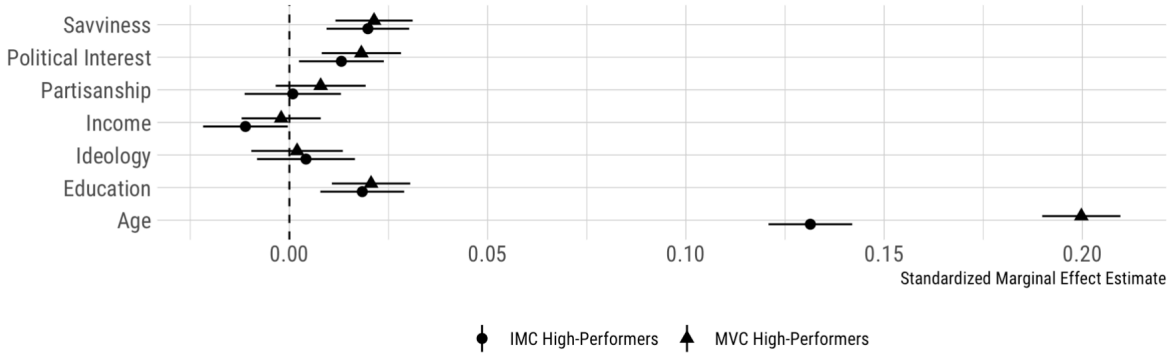


Notes: 95% CIs shown. Dependent variable is a binary indicator of passing (=1) vs. not passing (=0) a post-outcome factual manipulation check (FMC).

We next examine the relationship between the different attentiveness measures and passage of factual manipulation checks (FMCs) using linear probability models. Per Figure I3, moving from the minimum to maximum value of the MVC scale, we find that the probability of passing an FMC increases by 36 percentage points (pp). Moving from the minimum to maximum value of the IMC scale, FMC passage increases by only 28pp. This difference is statistically significant (Δ =.08; SE = .01; $p < .05$). Thus, MVCs also outperform IMCs in predicting post-treatment manipulation check measures that capture attentiveness to experimental stimuli.

Finally, we assess the demographic predictors of MVC and IMC high-performers. High-performers are defined as respondents who correctly answered at least two items within each scale. Per Figure I4, we find that savviness (measured as the number of prior surveys taken in the past four weeks; "Zero"=1; "More than 20"=5)), political interest, education, and age predict passage for both attentiveness measures. The estimates are remarkably comparable. Only in the case of age do we observe a statistically significant difference. Increasing age by one standard deviation increases MVC passage by 20pp, compared to 13pp for IMC passage. This difference is statistically significant (Δ =.07; SE = .007; $p < .05$). Overall, we find demographic comparability between MVC and IMC passers with the exception of age.

FIGURE I4. Demographic & Political Predictors of MVC/IMC Performance



Notes: 95% CIs shown. Outcome measures are a binary indicator of high performance on the IMC and MVC scales (<2 correct checks=0, >=2 checks=1).

In sum, we find that both MVCs and IMCs are useful tools for gauging attention.[7] Both measures predict larger treatment effects, more time spent on stimuli and the overall study, and factual manipulation check passage. In addition, those who score highly on both of these measures tend to have similar demographic characteristics, with the exception of age. However, we find that MVCs tend to possess modest but detectable advantages over IMCs in predicting slightly larger conditional average treatment effects, time spent on treatment stimuli and outcome variables, and better performance on factual manipulation checks that follow experimental vignettes.

[7] Indeed, we found the pairwise correlation (r) between IMC and MVC scales to be .51 ($p < .001$).

APPENDIX J:
COMPARING THE MOCK VIGNETTE AND 2SLS APPROACH

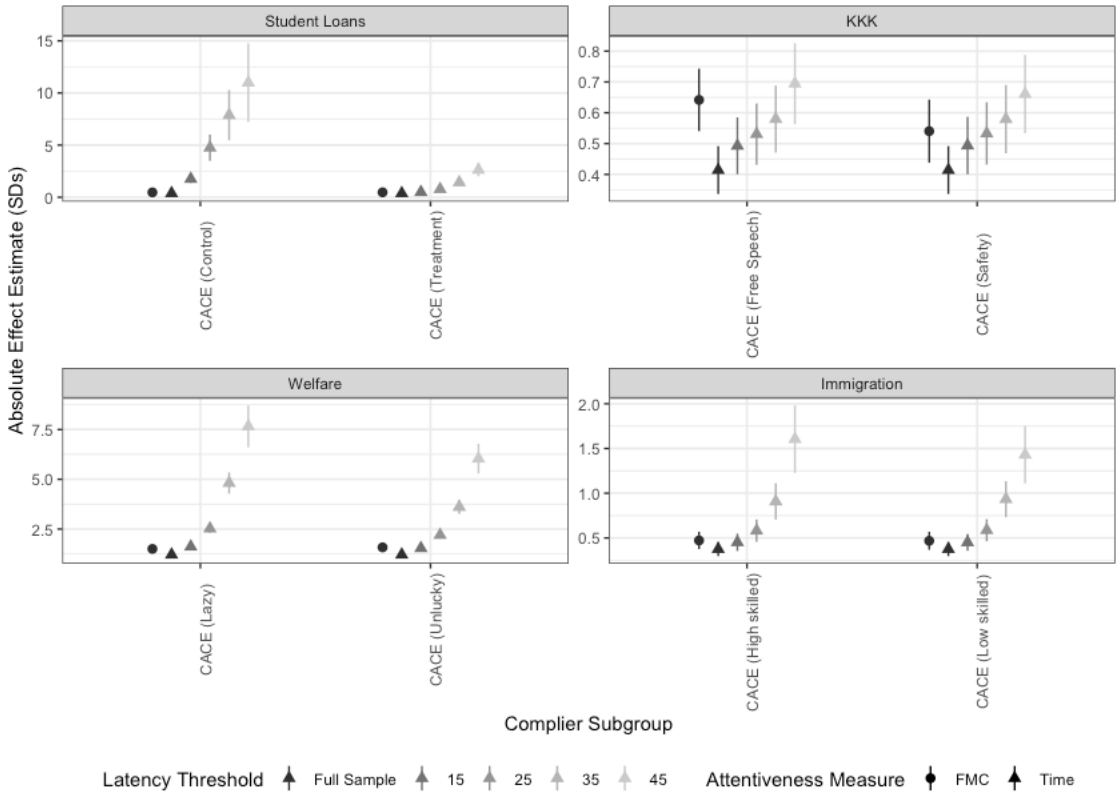
One approach to dealing with inattentiveness in experiments, though uncommon in practice, has been the use of two-stage least squares (2SLS) regression (e.g., see Harden, Sokhey, and Runge (2019)). However, it is important to note that results from such models are more difficult to interpret (Montgomery, 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. Regarding this latter point, proper implementation of the 2SLS method becomes even 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).

As an illustration of these points, Figure J1 displays complier average causal effect estimates (i.e., 2SLS estimates) for every experiment and condition in the Lucid 1 study reported in the manuscript. There are two different attentiveness measures featured: (1) response latencies (i.e., a timer on vignette to which one is assigned), and (2) responses to the factual manipulation check (FMC) that appeared after the control or treatment condition. For the response latencies, we feature various cut-off times in defining compliance. This enables us to create a binary measure from the original (continuous) latency measure, wherein 1=assigned to the treatment group *and* had a sufficiently high latency time, and 0=was assigned to the control group *or* did not have a sufficiently high latency time.

Following Harden, Sokhey, and Runge (2019), we designate one group as the non-compliant baseline group. Specifically, each member of this group is assigned a 0—either for the binary latency cut-off measure, or for passage of the (factual) manipulation check. In the designated treatment group, respondents are assigned a 1 when they either (1) spent more time on their respective experimental vignette than the designated cut-off, or (2) correctly answered the manipulation check that appeared after the outcome measure. (Otherwise, respondents in the treatment group are assigned a 0 for the attentiveness measure.)

FIGURE J1. Complier Average Causal Effects (CACEs) Across Experiments, Measures of Compliance, and Possible Latency Cut-offs



Thus, for each experiment, there are numerous CACEs one could potentially obtain: one per each experimental condition depending on which condition is designated as the treatment (e.g., in the KKK study, one could reasonably designate either condition as the "treatment"); different CACEs depending upon whether one uses a latency measure or a factual manipulation check to measure attentiveness; and, different CACEs depending upon the cut-off time that is used to construct the binary latency measure.

Such an array of modeling choices has the potential to yield substantively different findings. This challenge is illustrated in Figure J1. The first panel of this figure (Student Loan experiment) displays CACEs across different measures and thresholds of compliance with compliance defined by attentiveness to the control or treatment conditions. The second panel (the KKK experiment) presents CACEs with compliance defined by attentiveness to the free speech or public safety conditions. The third panel (the Welfare experiment) presents CACEs for the lazy and unlucky condition. Finally, the fourth panel (the Immigration experiment) displays CACEs for the low-skilled and high-skilled immigrant conditions. The black triangle represents the full sample treatment effect (i.e., ITT) for each condition. Absolute values of CACEs are computed to facilitate comparisons of effect size across conditions.

As shown in Figure J1, CACEs vary by condition, attentiveness measure, and compliance threshold. Depending on how a researcher defines compliance and which group they designate as the treatment group, CACEs range from .46 to 10.99 standard deviations in the Student Loans experiment, .49 to .69 standard deviations in the KKK experiment, 1.5 standard deviations to 7.6

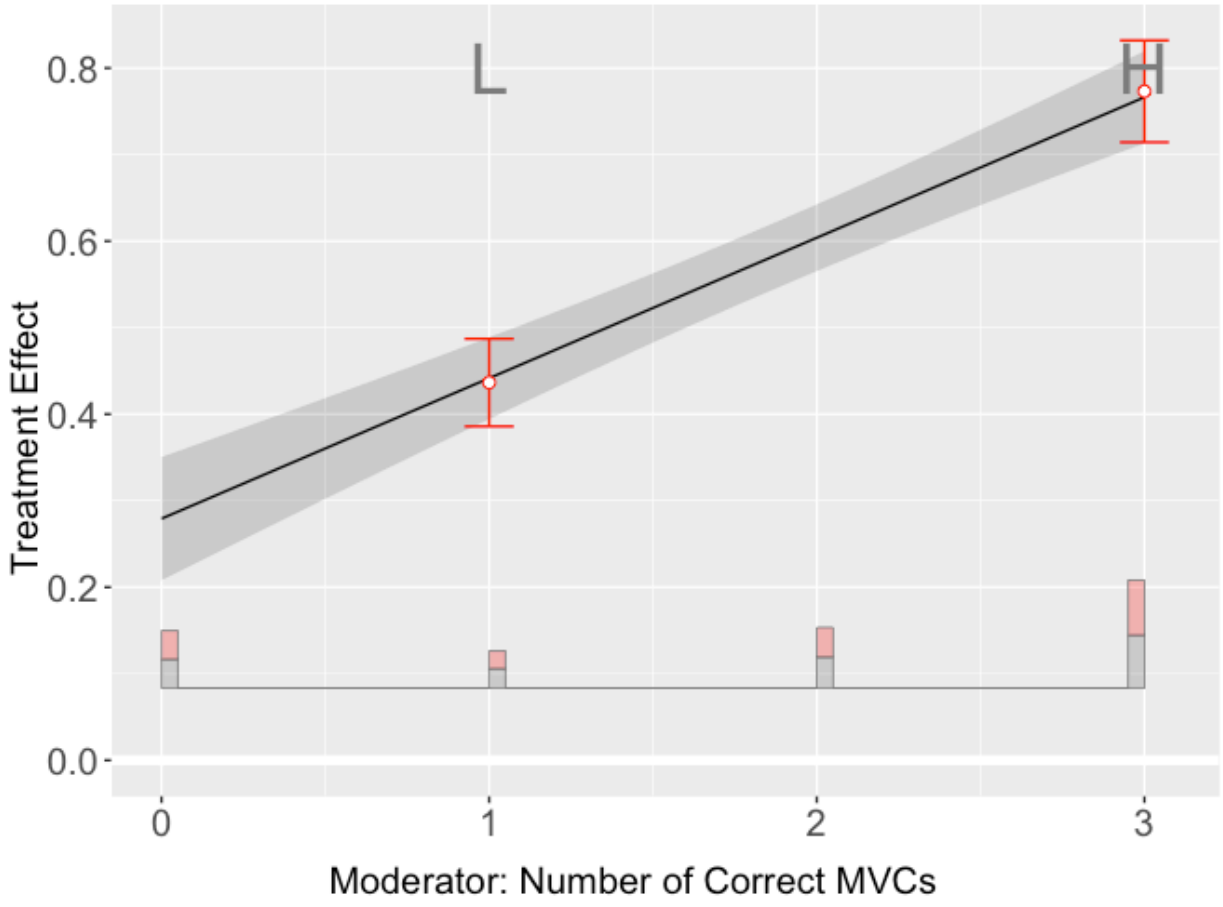
standard deviations in the Welfare experiment, and .45 to 1.6 standard deviations in the Immigration experiment. Thus, there is a considerable variation in CACEs even within the same experiment. Moreover, CACEs can take on implausible values depending on which compliance thresholds are used. This is not surprising, given that the CACE is equivalent to the ITT divided by the share of compliers in the sample (i.e., the Wald estimator). Higher thresholds for compliance decrease the proportion of compliers, magnifying the CACE as a result. For example, a one standard deviation ITT can produce a CACE of 2.5 standard deviations if only 40% of the treatment group complied with the treatment.

Thus, while the 2SLS approach is a reasonable method for identifying treatment effects among attentive respondents, its implementation, and the interpretability of results, are relatively more complex vis-à-vis the Mock Vignette technique we propose.

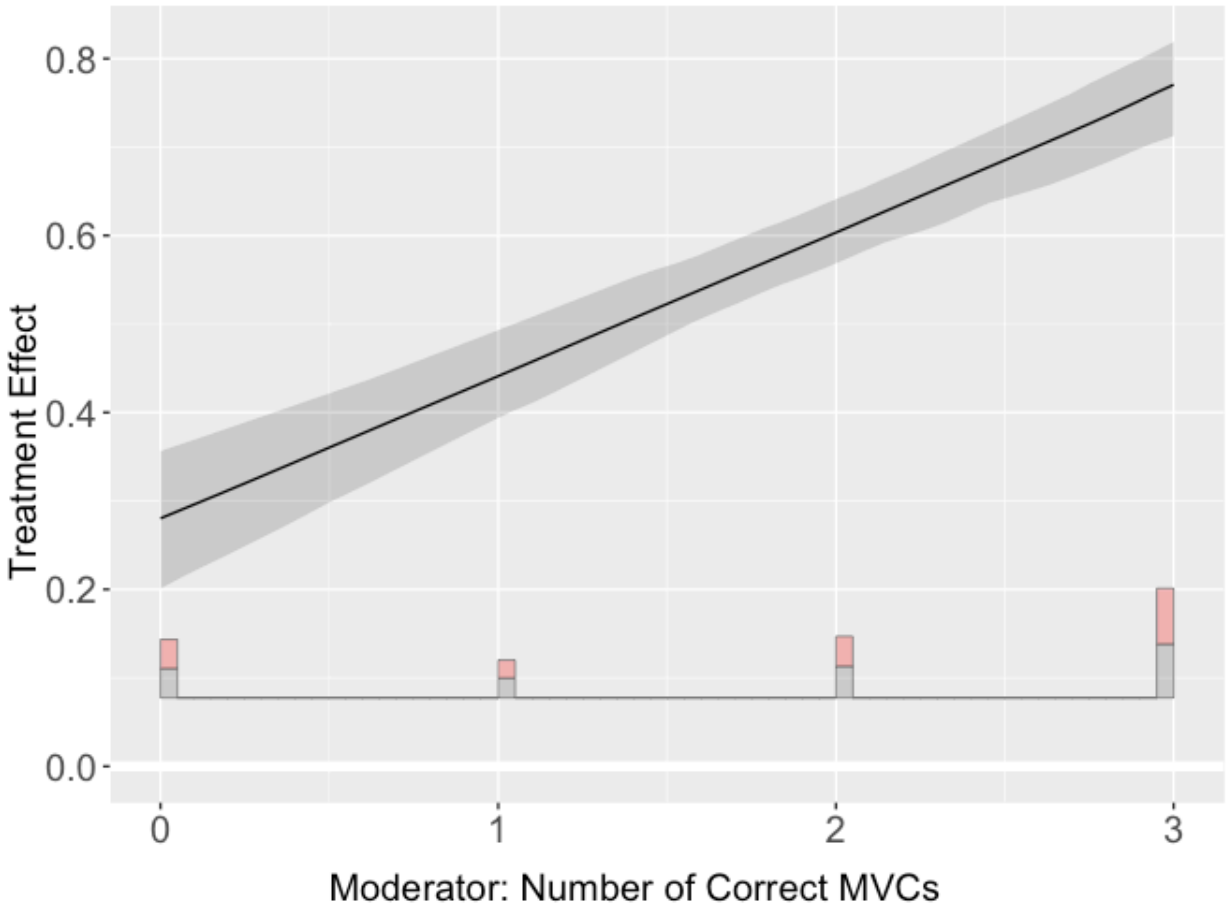
SUPPLEMENTAL APPENDIX K:
TESTING THE LINEAR INTERACTION ASSUMPTION

Using Hainmueller, Mummolo, and Xu (2019)'s binning estimator, we find that, using our Lucid data set, the linear multiplicative model is largely consistent with the data (see Figure K1a). The *p*-value of the Wald test is .488, and thus, we fail to reject the null hypothesis that the linear multiplicative model and the two-bin model are statistically equivalent. Estimation of our interaction model using a flexible kernel smoothing estimator corroborates this analysis, as we can see that the functional form is approximately linear (see Figure K1b). Thus, we find that specifying a linear interaction (i.e., treating the moderating variable (the MVC performance scale) as continuous) is justifiable.

FIGURE K1: Binning And Kernel Estimates Of Conditional Effects



(a)


0.8
0.6
0.4
0.2
0.0
Treatment Effect
0
1
2
3
Moderator: Number of Correct MVCs

(b)