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, .79]
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, .72]
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) for MVCs 1, 2 and 3. 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	11.18	15.51	27.18	46.13
Treatment	10.60	15.38	26.98	47.03
Overall	10.97	15.50	27.04	46.49

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. The Lucid results are for MVCs appearing in the second round of the study.

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.

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.

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 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=5,890)	Lucid 2 (N=9,148)
Median Income	50k-75k	25k-50k	50k-60k	50k-75k	25k-50k	25k-50k
Median Education.	College Grad	Some College	Some College	College Grad	Some College	Some College
Mean Age	36.82	46.56	48.57	38.28	48.50	48.41
Female	51.74%	50.00%	51.21%	50.62%	45.09%	51.22%
White	71.48%	62.37%	68.15%	74.38%	74.69%	74.08%
Black	7.63%	11.86%	10.75%	8.58%	10.53%	11.88%
Hispanic	11.77%	17.09%	14.65%	8.33%	7.08%	12.57%
Democrat	52.07%	46.68%	45.81%	52.86%	43.72%	44.39%
Independent	15.26%	19.26%	17.16%	17.04%	18.85%	20.40%
Republican	32.67%	34.06%	37.03%	30.10%	37.44%	35.21%
Liberal	49.92%	35.97%	-	47.76%	32.84%	32.37%
Moderate	15.75%	28.95%	-	21.14%	32.39%	33.29%
Conservative	34.33%	35.08%	-	31.09%	34.77%	34.35%
Mean Political Interest	3.50	3.44	-	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).

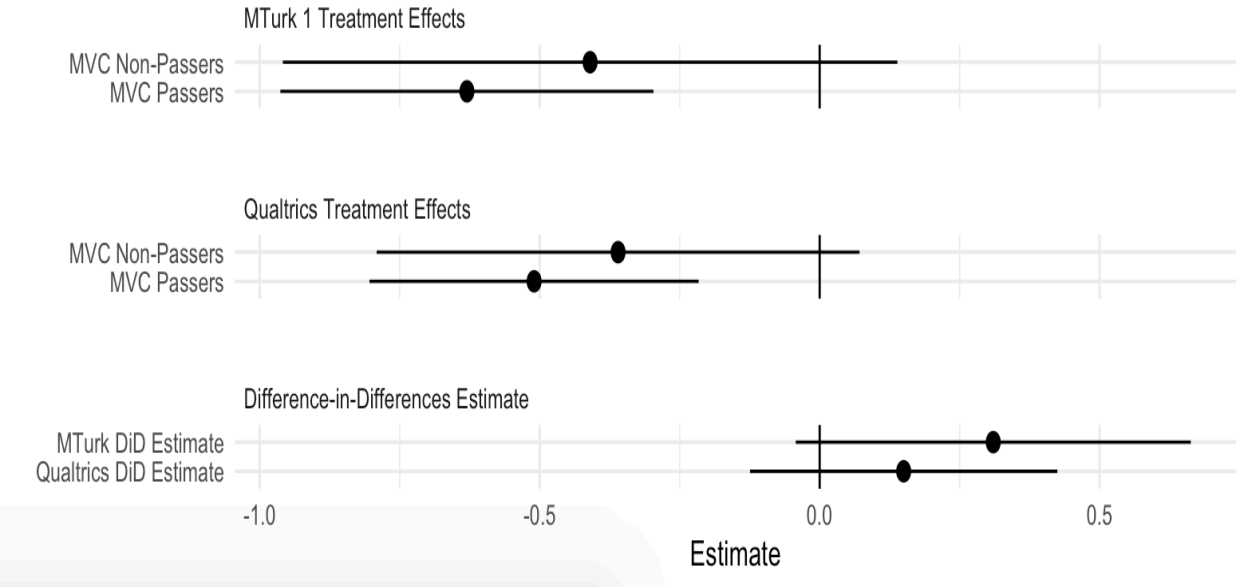
APPENDIX C:
RESULTS FOR MTURK 1 & QUALTRICS STUDIES

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 D1). 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 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***	-0.04
	(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	-0.04*
	(0.09)	(0.08)	(0.05)	(0.05)	(0.02)
Education	-0.11	0.11	0.27***	-0.12†	0.05*
	(0.11)	(0.09)	(0.08)	(0.07)	(0.02)
Political Interest	0.02	0.08	--	-0.06	0.07***
	(0.07)	(0.06)		(0.05)	(0.02)
Party ID	-0.11	-0.02	-0.01	-0.02	-0.06**
	(0.09)	(0.06)	(0.04)	(0.05)	(0.02)
Ideology	-0.15†	-0.09	--	-0.08	0.01
	(0.09)	(0.07)		(0.05)	(0.02)
Constant	0.75***	0.43***	0.31***	0.66***	0.45***
	(0.08)	(0.07)	(0.06)	(0.05)	(0.02)
N	603	784	742	804	9,900
R-squared	0.11	0.07	0.09	0.11	0.12

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%*	.33*
KKK Experiment (Qualtrics)	83%*	88%*	81%*	12%*	10%*	.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=1358)	200%*	101%*	158%*	32%*	63%*	.40*
KKK Experiment (n=1362)	172%*	170%*	162%*	34%*	64%*	.50*
Welfare Experiment (n=1359)	203%*	157%*	155%*	67%*	65%*	.43*
Immigration Experiment (n=1356)	205%*	168%*	164%*	63%*	70%*	.48*

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*
	(.039)
Treatment × MVC Score × Scientific Publishing MV	-.020
	(.056)
Treatment × MVC Score × Stadium Licenses MV	-.008
	(.054)
Treatment × MVC Score × Sulfur Dioxide MV	-.022
	(.054)
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	4,286

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*=1,179) 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.

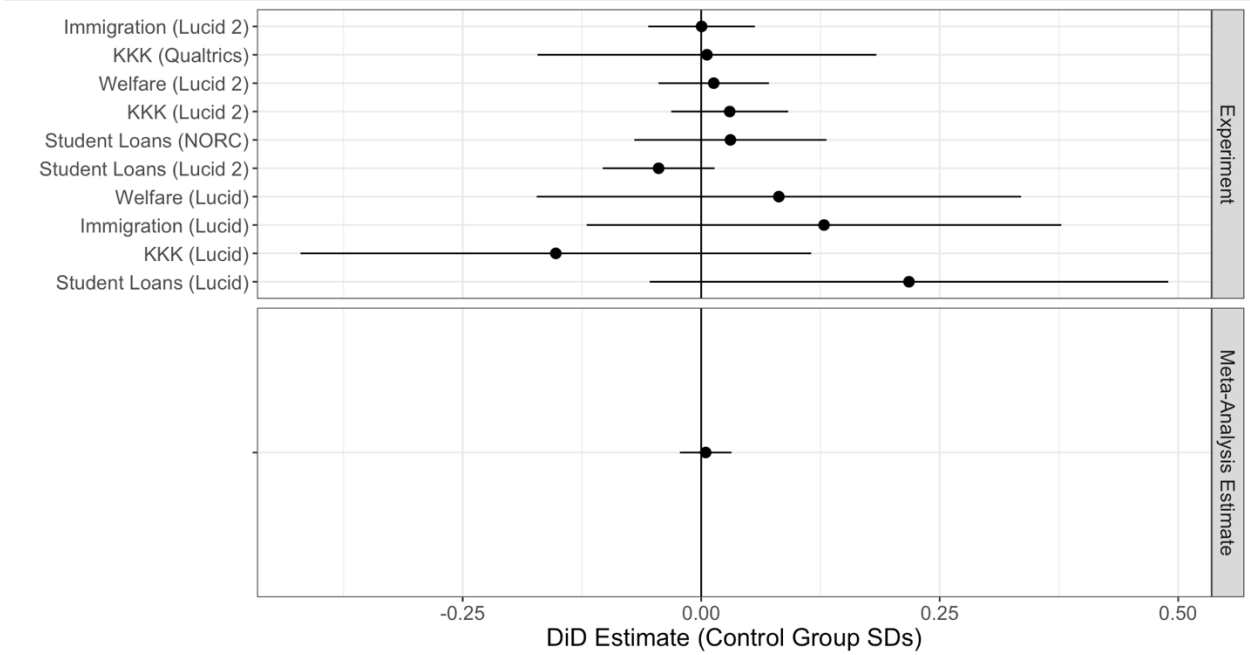
As demonstrated in Figure 4, we find no evidence that those respondents who observed, versus did not observe, an MV before the experiment exhibited significantly different treatment effects. Treatment effects were, in each experiment, substantively and statistically similar across these two groups. Indeed, 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.

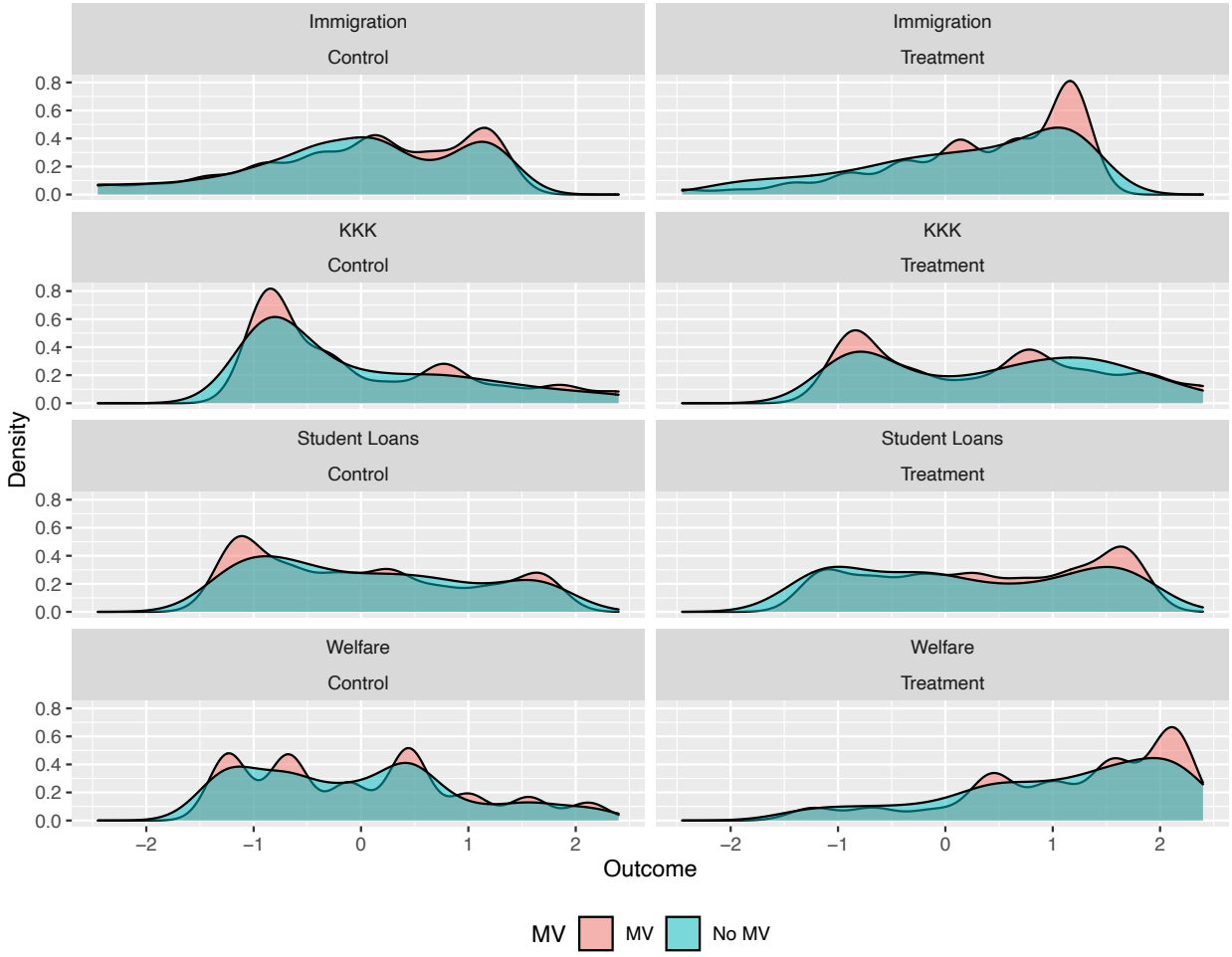
[4] In contrast to fixed-effect meta-analysis, which assumes that studies are estimating a single "true" effect, random-effects meta-analysis models assume that effects are drawn from a larger population, and may vary from study to study. In our case, the fixed-effect meta-analysis estimate (.0047) is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

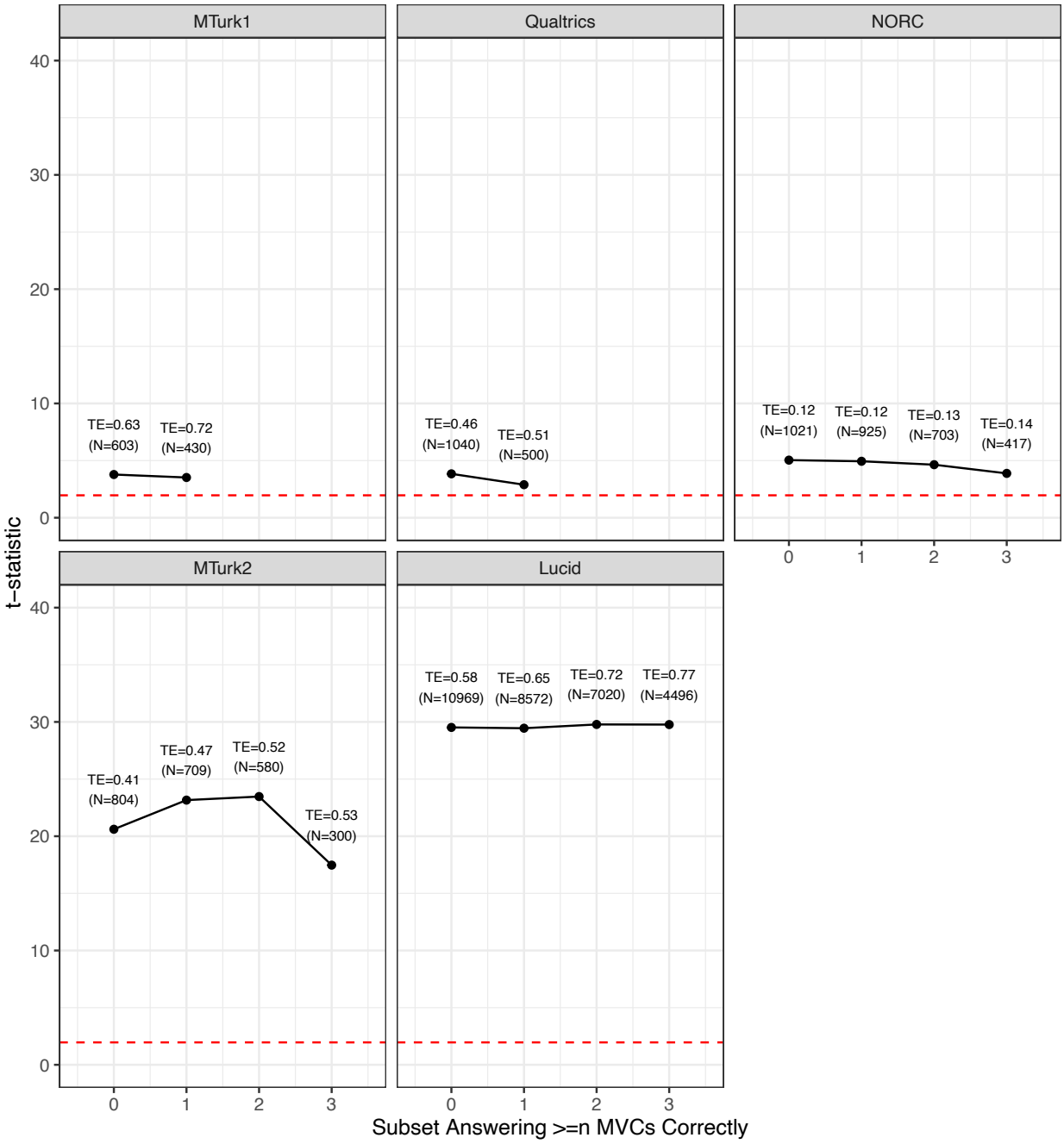
As *t* is a function of the estimated treatment effect (TE) divided by the standard error (SE), change in this statistic is a useful indicator of the net consequence (on our ability to detect a statistically significant effect) of 1) a larger effect, yet 2) smaller sub-sample.

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

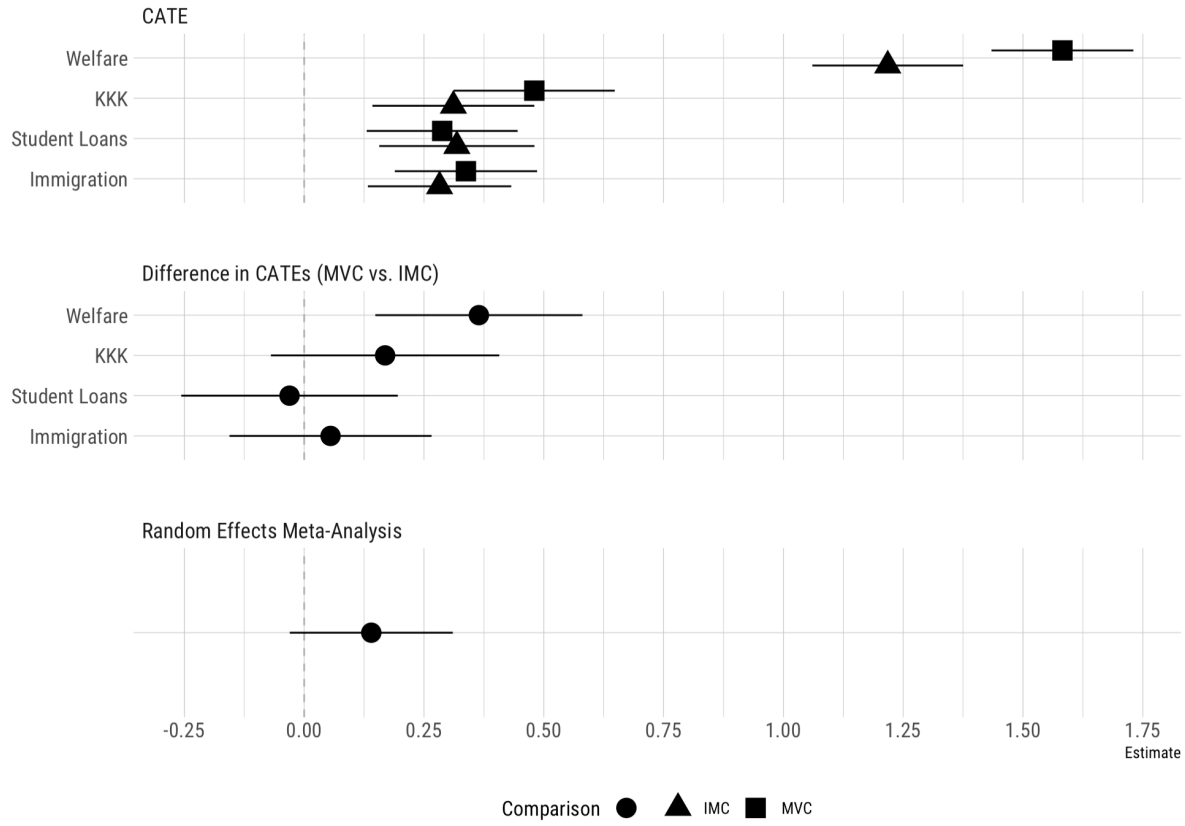
We begin with comparing MVC and IMC CATEs. Table I1 shows that pooling across experiments, the difference in treatment effects between the least and most attentive, as measured by the MVC, is .69 control group standard deviations (SE = .04; p < .05). For the IMC measure, the difference in treatment effects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

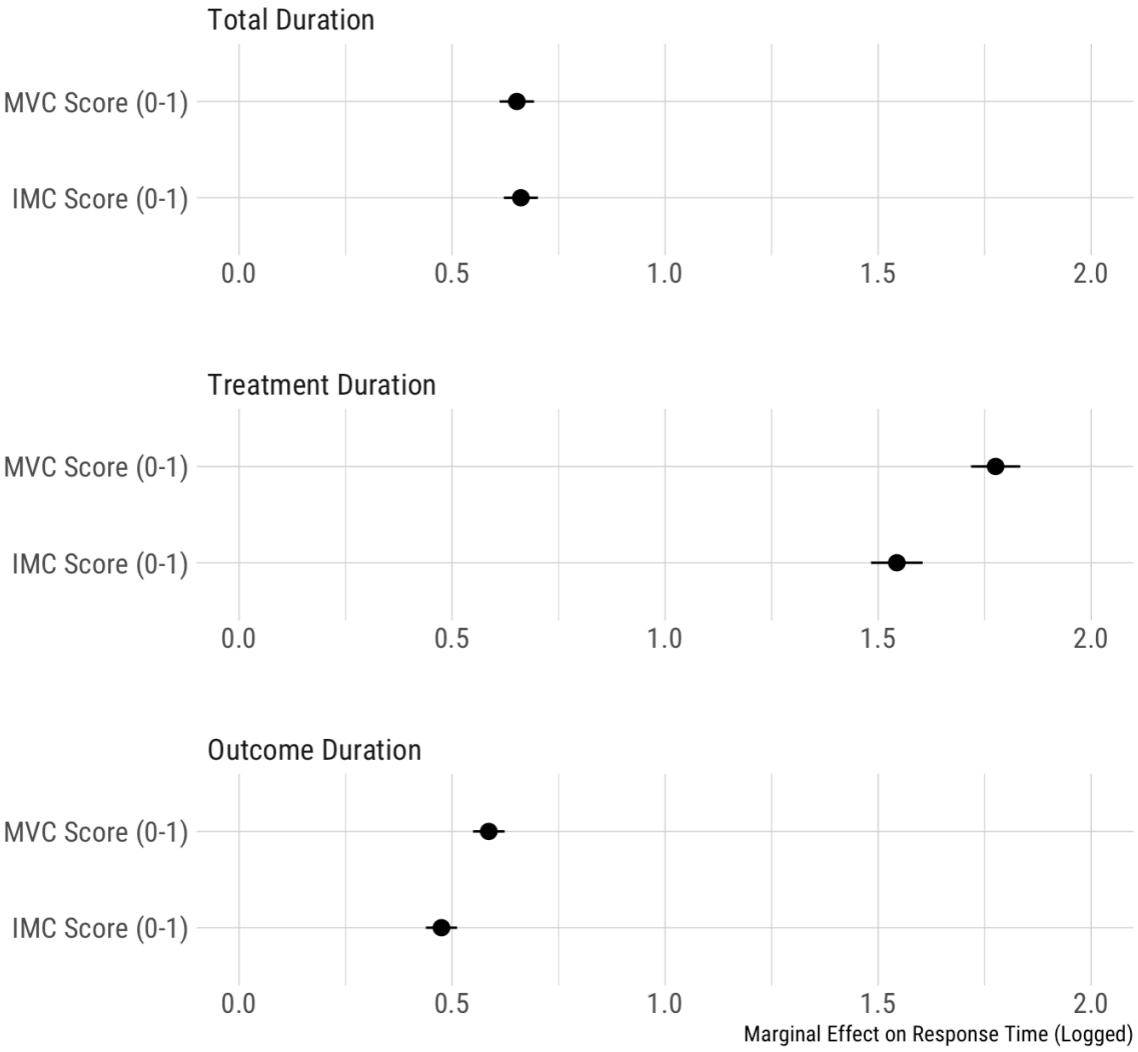
FIGURE I1. MVC Performance, IMC Performance, and CATE Estimates (By Experiment)



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

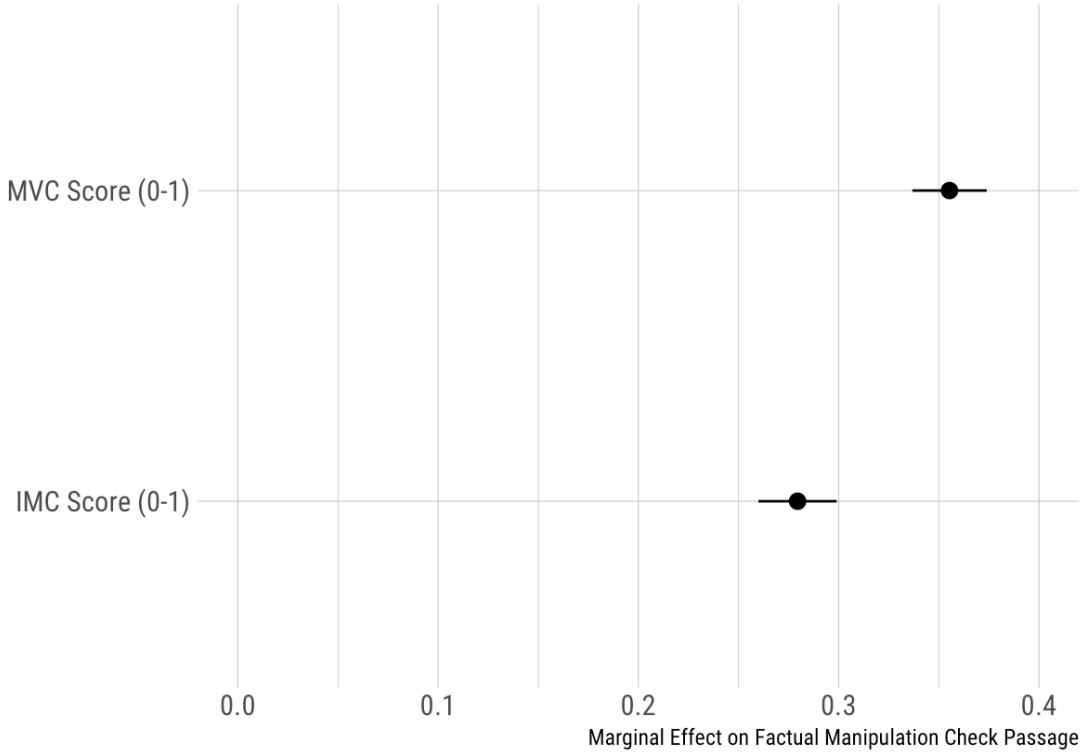


Notes: 95% CIs shown. All outcome measures are log-transformed.

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



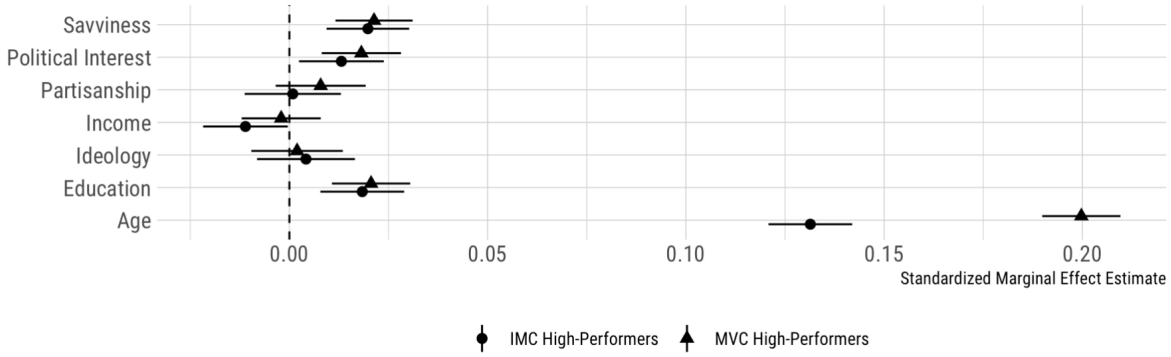
Notes: 95% CIs shown. Dependent variable is a binary indicator of passing (=1) vs. not passing (=0) a post-outcome factual manipulation check (FMC).

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 ($\Delta = .07$; 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

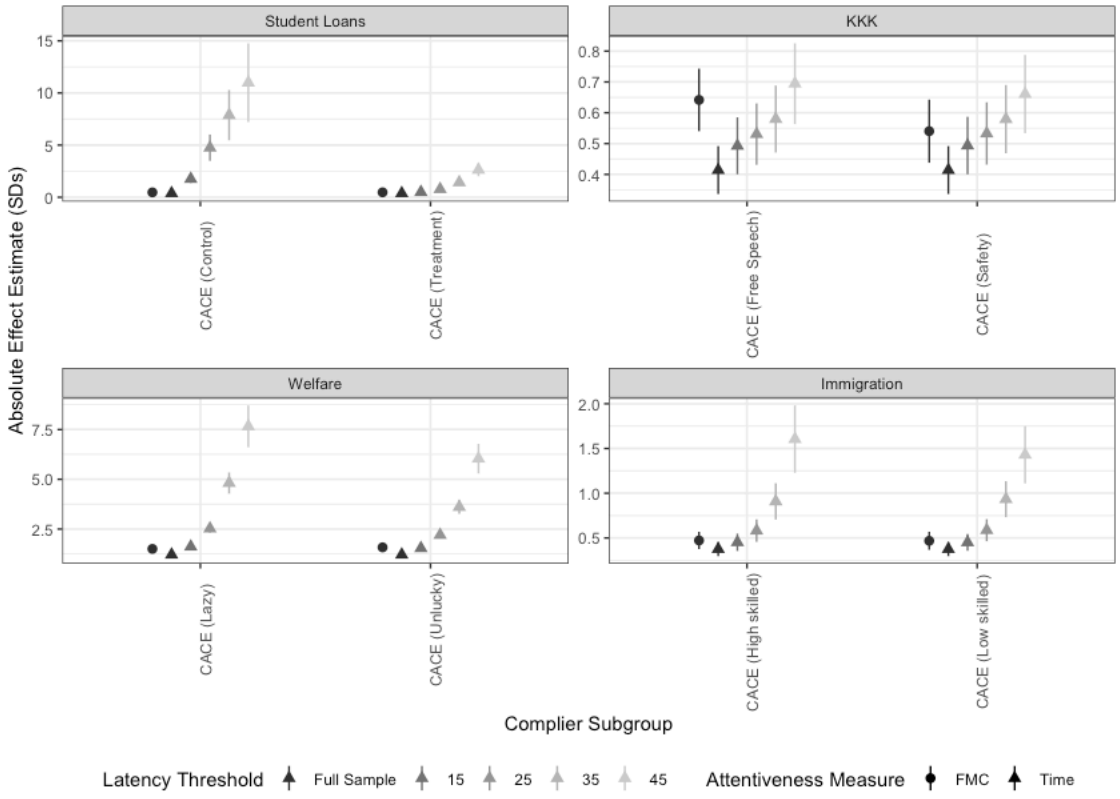
One approach to dealing with inattentiveness in experiments, though uncommon in practice, has been the use of two-stage least squares (2SLS) regression (e.g., see Harden, Sokhey, and Runge (2019)). However, it is important to note that results from such models are more difficult to interpret (Montgomery,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

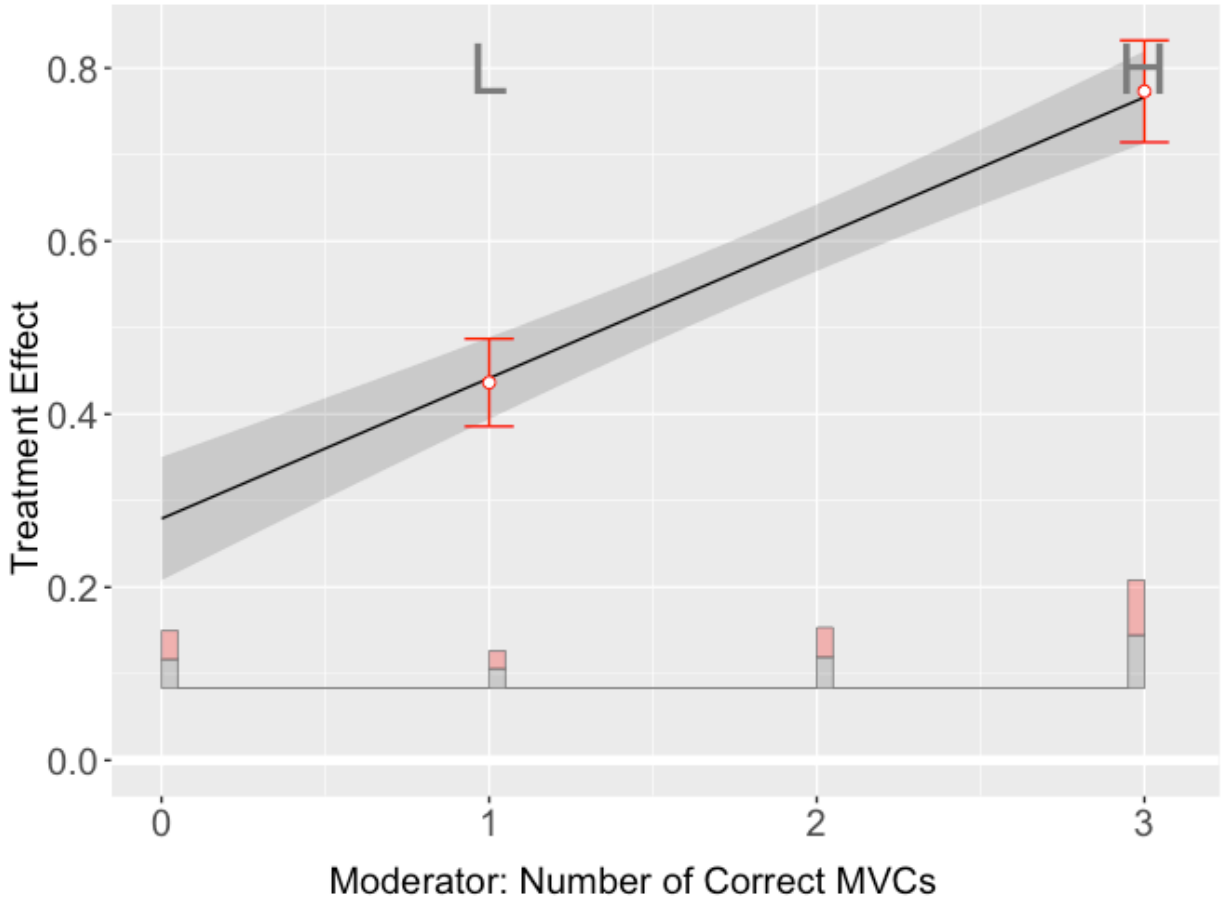
standard deviations in the Welfare experiment, and .45 to 1.6 standard deviations in the Immigration experiment. Thus, there is a considerable variation in CACEs even within the same experiment. Moreover, CACEs can take on implausible values depending on which compliance thresholds are used. This is 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.

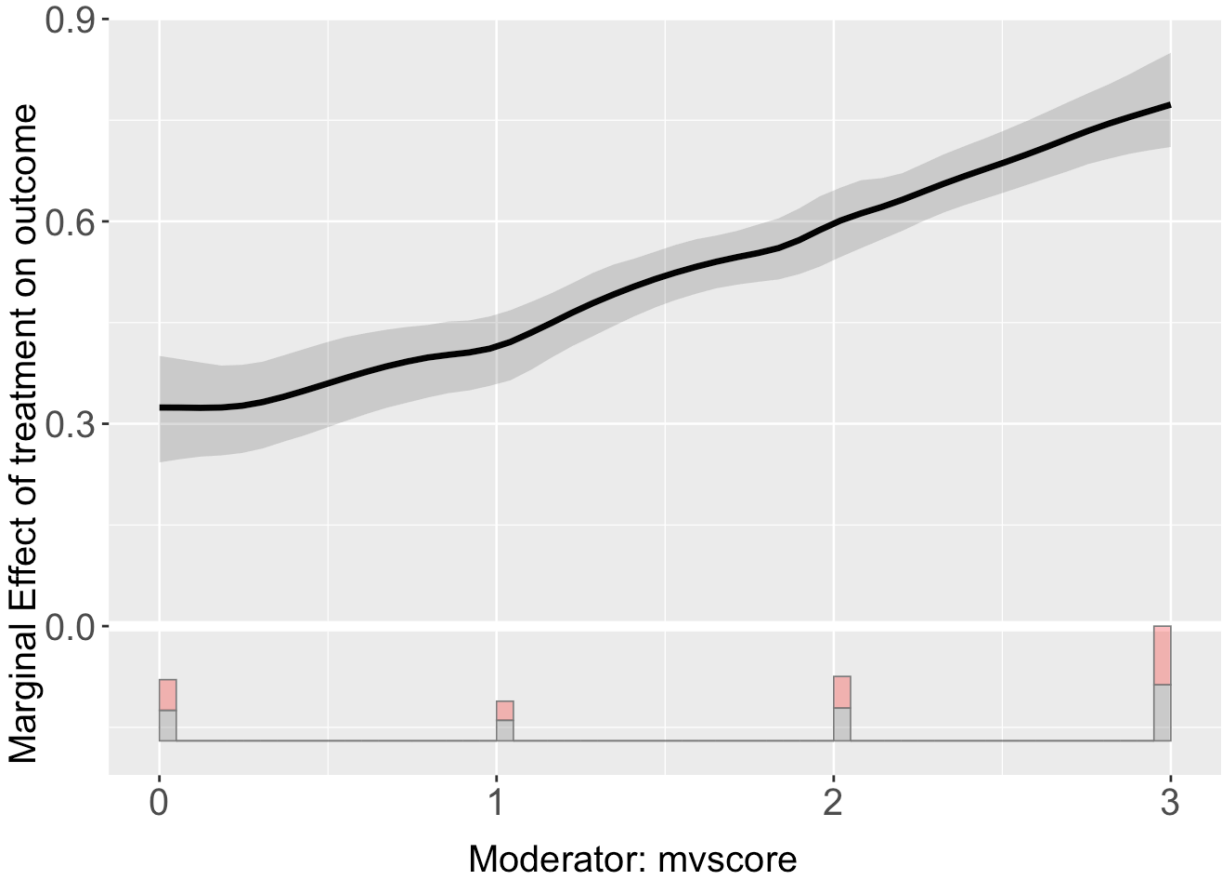
SUPPLEMENTAL APPENDIX K:
TESTING THE LINEAR INTERACTION ASSUMPTION

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)


0.9
0.6
0.3
0.0
Marginal Effect of treatment on outcome
0
1
2
3
Moderator: mvscore

(b)