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)

Presented at the International Methods Colloquium
Friday November 6, 2020

Survey Experiments

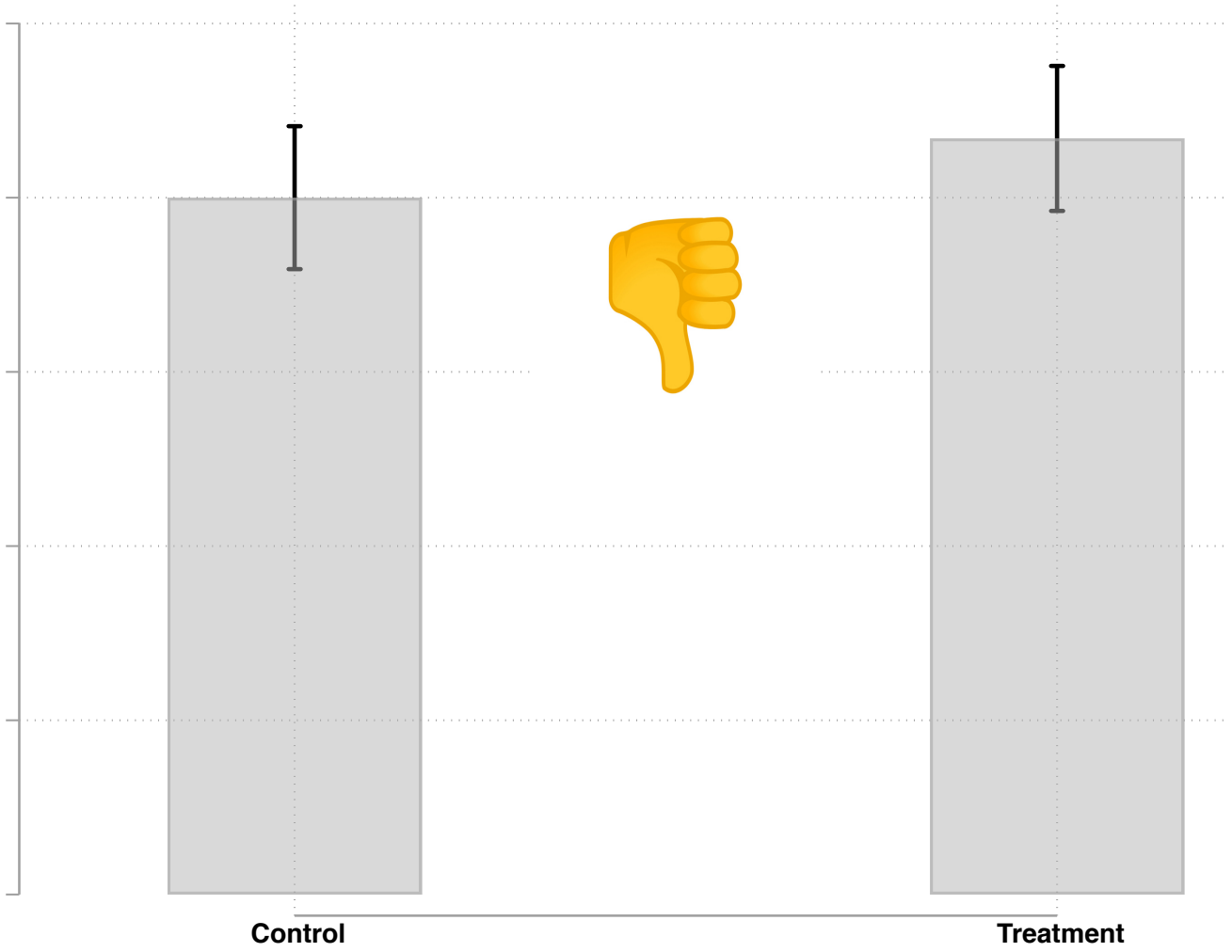
- Popular method for testing hypotheses
- Increasingly conducted via online platforms (MTurk, Qualtrics, Lucid) rather than in lab settings
- Large, adult samples at low cost
- Utility of a survey experiment hinges upon attentiveness to content (i.e., compliance)

Challenge #1: Inattentiveness as Noncompliance

- A major concern: are respondents *actually* being attentive?
  - Existing research finds inattentiveness to be substantial
- The statistical stakes: inattentiveness reduces estimated treatment effects (in expectation)
  - Gerber and Green (2012)
- The practical stakes: Null or weak effects could be due to inattentiveness, not bad theory

A Familiar Situation

- We go through the hard work of setting up a survey experiment (designing, IRB submission & modifications, programming, fielding, compensating, answering angry respondent emails…)
- Result? Null, or very weak, effects
- Why did this happen?
  - Theory is lousy?
  - Poor operationalization of the independent variable?
  - Inattentiveness?
  - All of the above?
- We often don't know, and might even abandon the project rather than risking more time/$/effort



What to do?

- Adjudicate between these possibilities with an individual-level measure of attentiveness to the experiment

- Existing methods:

  - Timers (aka, latency measures) on survey screens (Niessen, Meijer, and Tendeiro 2016; Wood et al. 2017)
    - Shorter generally indicates less attention

  - Instructional manipulation checks (IMCs/Screeners) (Berinsky, Margolis, and Sances 2014; Oppenheimer, Meyvis, and Davidenko 2009)
    - Trick questions involving content unrelated to the experiment (and unrelated to each other)

  - Factual manipulation checks (FMCs) (Kane and Barabas 2019)
    - Factual questions (with right or wrong answers) about experimental content

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

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

John V. Kane New York University
Jason Barabas Stony Brook University

- Few studies actually employ manipulation checks (18% between 2001-2015), especially to check for attentiveness

- Factual manipulation checks (FMCs) have useful diagnostic functions:
  - Determine extent to which inattentiveness is a problem
  - Whether assignment significantly correlates with responses to FMCs
  - Whether some treatments were relatively difficult to perceive etc.

- But, researchers ALSO want to use manipulation check passage to re-estimate treatment effects
  - E.g., via subsetting, controlling, or interacting with treatment

Challenge #2: Post-Treatment Bias

- Measures of attentiveness may, themselves, be affected by treatment
- Conditioning on a post-treatment variable threatens to "de-randomize" experimental groups (Coppock 2019)
- The comparison may potentially be between "dissimilar groups" (Montgomery, Nyhan and Torres 2018)
  - Comparing those with high MC performance under one condition against those with high MC performance under a different condition
- If the covariate(s) responsible for T vs. C differences has a non-zero correlation with Y, we obtain a biased estimate

What to do now?

- The Ideal: Can we have a measure of attentiveness to one's experiment that is *not* post-treatment?

- By definition, no.
  - We can't directly gauge attentiveness to X *before* X is observed.

- Next best alternative: A measure of attentiveness…
  - To (generally) similar kind of content as one's experiment (often, a vignette)
  - Occurring immediately before one's actual experiment
    - Attentiveness fluctuates throughout surveys (Berinsky, Margolis, and Sances 2014)

- Enter: Mock Vignettes

Mock Vignettes (MVs) & Mock Vignette Checks (MVCs)

- MVs: Brief vignettes about (vaguely) policy-related content
  - Simulate the experiment itself (exposure to socio-political content, then follow-up questions)
  - Occur immediately before an experiment
  - Same MV viewed by all respondents

- MVCs: Follow-up questions about the MV that have only one correct answer
  - Appear on the following screen; can't return to previous screen to look up answer

Example

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 Checks (MVCs)

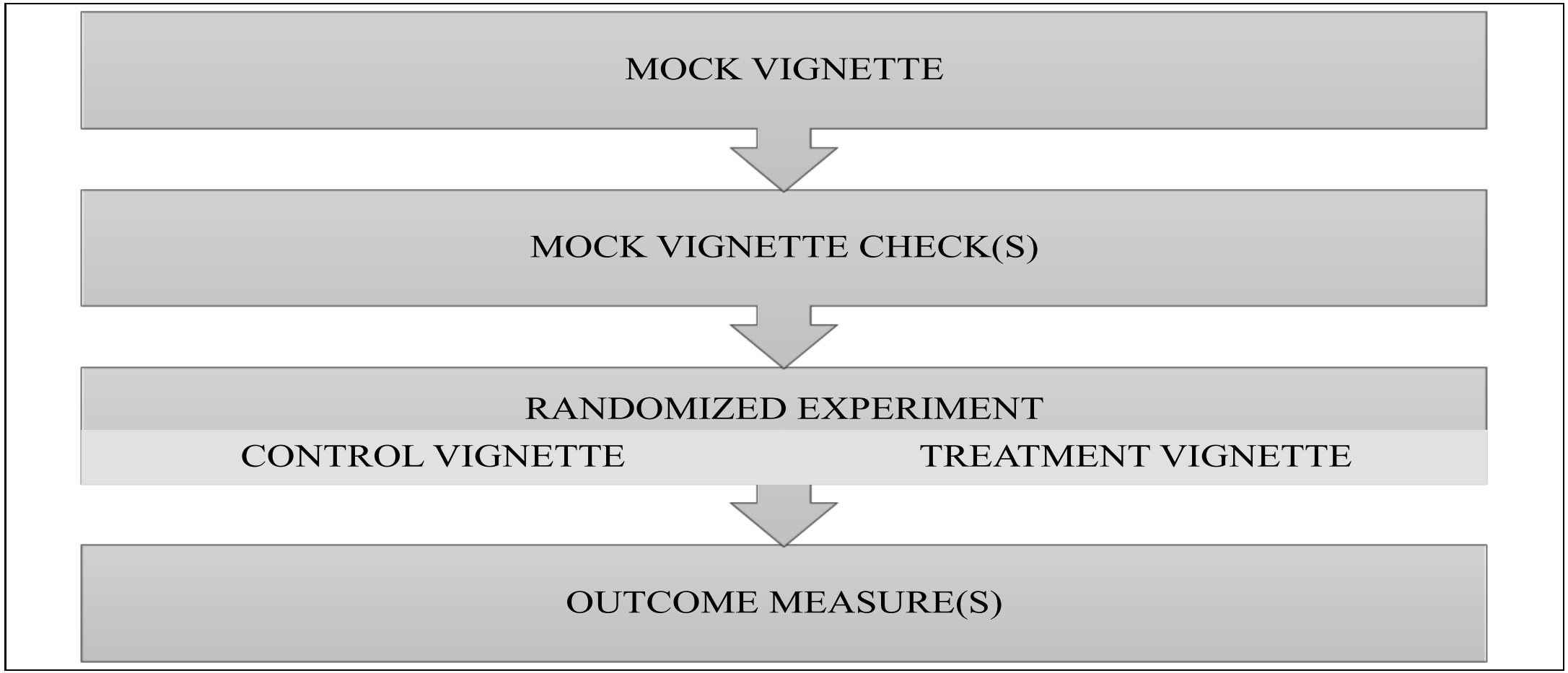
- MVCs enable researcher to create a pre-treatment, individual-level measure of attentiveness
- Serves as a proxy measure of attentiveness to the actual experimental vignette
- Because MVCs are pre-treatment, can be used to subset, interact, etc. without risking post-treatment bias
- Our primary research question: *does MVC performance predict larger treatment effects?*

Data and Methods

- **General design:** have participants read a MV, measure attention using MVCs, randomly assign participants to experimental conditions, and measure experimental outcomes.
  - In the Lucid design (n ~6,000), each respondent participated in 2 rounds

Data and Methods

FIGURE 1. Implementation of Mock Vignettes in Each Study



Data and Methods

- **General design:** have participants read a MV, measure attention using MVCs, randomly assign participants to conditions, and measure outcomes.
  - In the Lucid design (below), each respondent participated in 2 rounds

- **Data sources:** Qualtrics, MTurk, NORC, and **Lucid**
  - Variation in sampling method (probability vs. non-probability) and subject pool
  - 5 studies (total n >9,000); replicated 4 published experiments

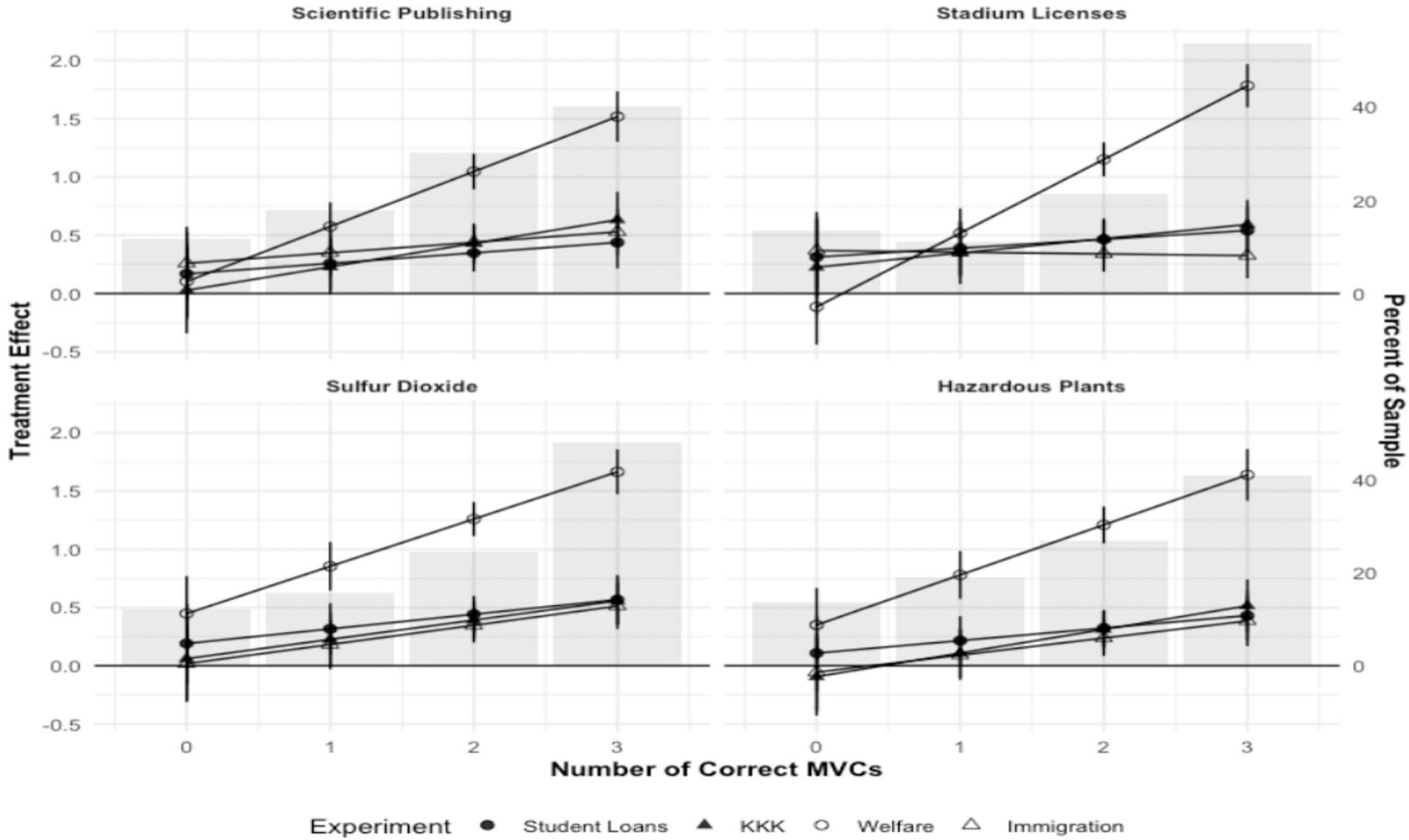
  - **Analysis:** Examine conditional average treatment effects (CATEs) using a linear model regressing outcomes on treatment indicator, MVC score (additive scale ranging 0 to 3), and their interaction

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	Event 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 Appendix. "Student Loan Forgiveness" = Mullinex, Leeper, Druckman and Freese (2015); "KKK Demonstration" = Nelson, Clawson and Oxley (1997); "Welfare Deservingness" = Aarøe and Peterson (2014); "Immigration Policy" = Valentino et al. (2019).

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



Summary of Key Findings

- MVC scores are associated with larger CATEs across multiple samples and experiments.

- CATEs are generally weak and non-significant among MVC non-passers, but statistically and substantively significant among MVC passers in every case

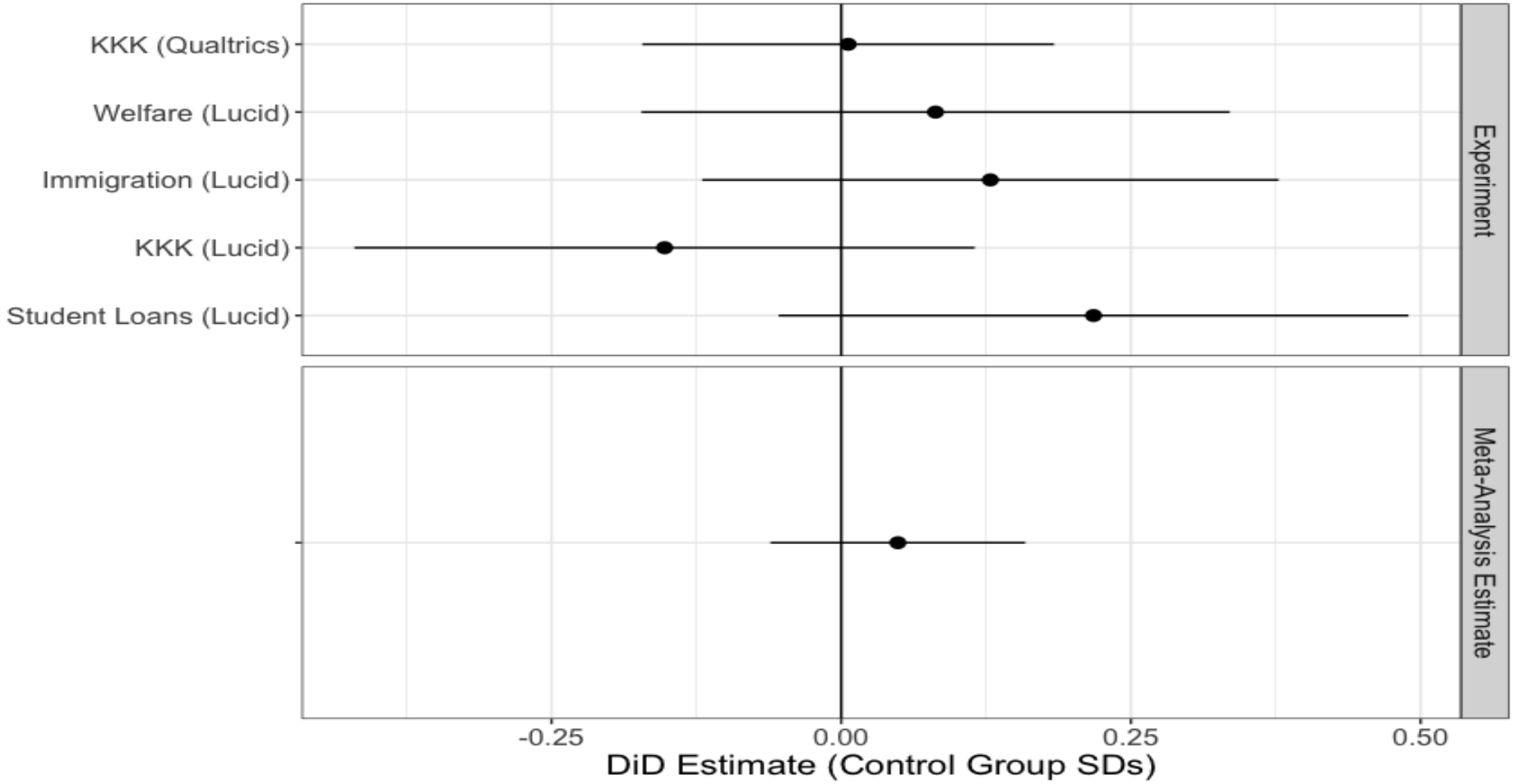
- MVC scores correlate with other indicators of attentiveness
  - Timers: In every study, higher MVC scores predict significantly more time spent on MVs, experimental vignettes, outcome measures, and surveys
  - FMCs: In every study, higher MVC scores predict significantly higher (factual) manipulation check passage rates (change in predicted probability as MVC scores move from min- to maximum ranges from .35 to .68)
  - Confirms that MVC performance is a reasonable proxy for attention to experiment

Does Using Mock Vignettes Alter Treatment Effects?

- Randomly assigned whether a MV/MVC appeared before an experiment (Qualtrics and Lucid studies)

- Compare treatment effects with vs. without an MV/MVC appearing before an experiment

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



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

Summary of Additional Findings & Diagnostics

- Lucid MVC passage rates ranged from 50-80%

- MVs are interchangeable and MVC performance strongly correlates across *rounds* (*r* =.60)

- MVC performance does not have significant *political* correlates
- A couple of consistent demographic correlates (e.g., age and race), but weak correlations (age *r* =.32; non-white *r* = -.17 (Lucid))
  - Further, sample composition barely changes when looking at high MVC performers
  - CATEs (and their *p*-values) remain nearly identical when specifying a controlled interaction with a correlate of MVC performance

- Linear multiplicative model (simpler) is justified (Hainmueller, Mummolo and Xu (2019))

Summary of Additional Findings & Diagnostics

- In our Lucid study, not a big difference in effects with MV shortly vs. immediately before experiment

- Test statistics (t) do not necessarily decline as we analyze more attentive sub-samples
  - The larger treatment effect helps to compensate for decrease in n

- We offer researchers a handful of pre-tested MVs/MVCs
  - Include passage rates, reading time, measure of complexity, and IRT analyses of MVC difficulty and discrimination

Discussion & Conclusion

- Mock Vignette Checks (MVCs): a tool for gauging attentiveness that 1) serves as a proxy for respondent attention during the experiment, and 2) does not risk post-treatment bias

- MVC performance predicts larger CATEs and correlates with other measures of attention

- Examining whether effects are *stronger* at higher MVC performance allows for a hypothesis test more robust to inattentiveness
  - Always report intention-to-treat (ITT) effect first for full transparency
  - A stronger effect among attentive would constitute stronger evidence for a hypothesis
  - If treatment effect is weak and non-significant for sample as a whole and among the more attentive, problem is likely less about inattentiveness and more with theory or operationalization of IV (investigate the latter with a MC)

Discussion & Conclusion

- Ideal for typical (i.e., vignette-based) survey experiments, but still potentially useful for other types of experiments requiring attentiveness
- Other techniques (e.g., IMCs, IV/CACE) not mutually exclusive
- Though no measure of attentiveness is perfect, better than not accounting for inattentiveness at all
- Get the most out of your experimental data
    - Don't let inattentiveness ruin your treatment effects

Thank You!

IMC_MVC_1. This presentation was about:

A. The Electoral College
B. COVID-19
C. Mock Vignettes
D. Funny TikTok Videos

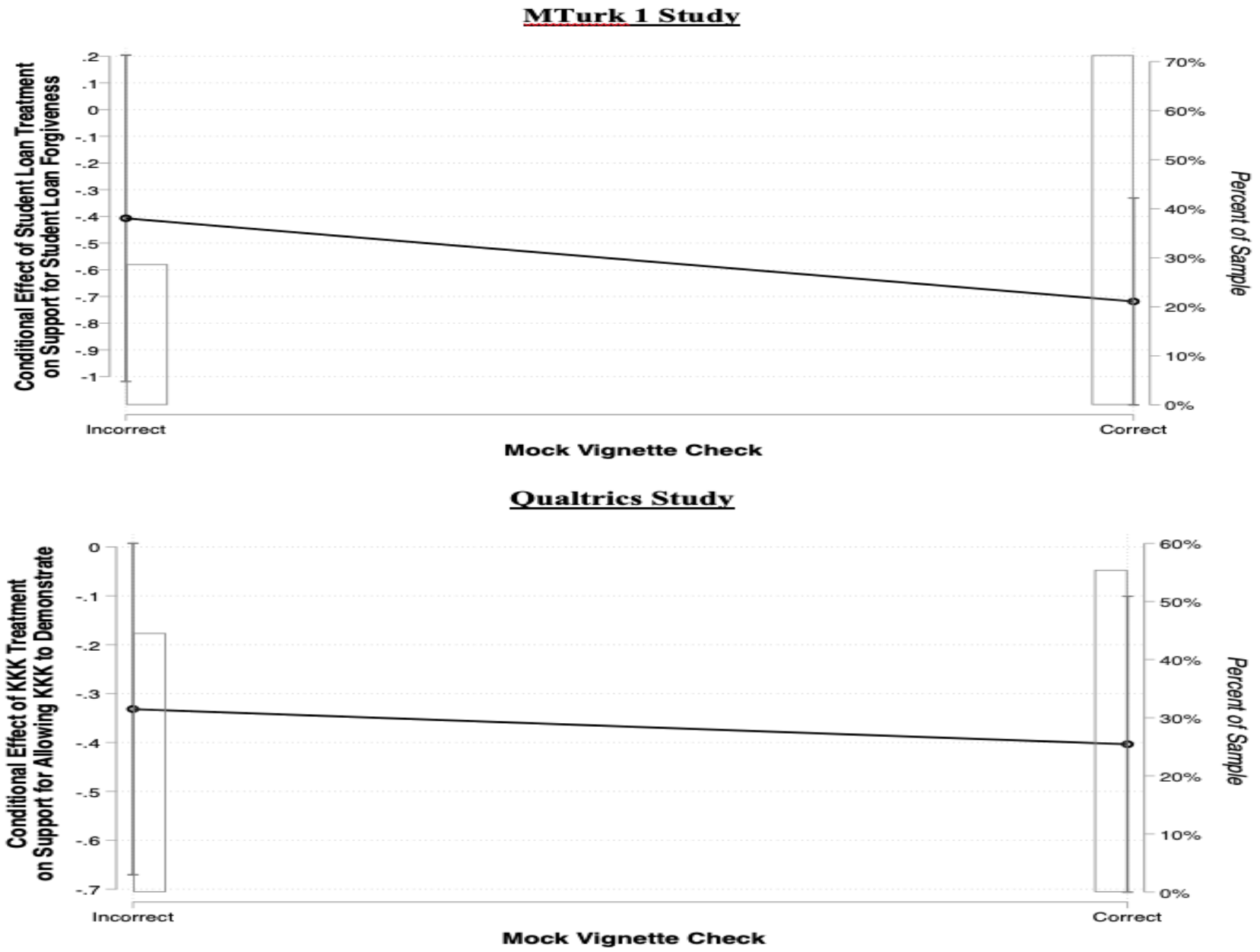
Supplemental Slides

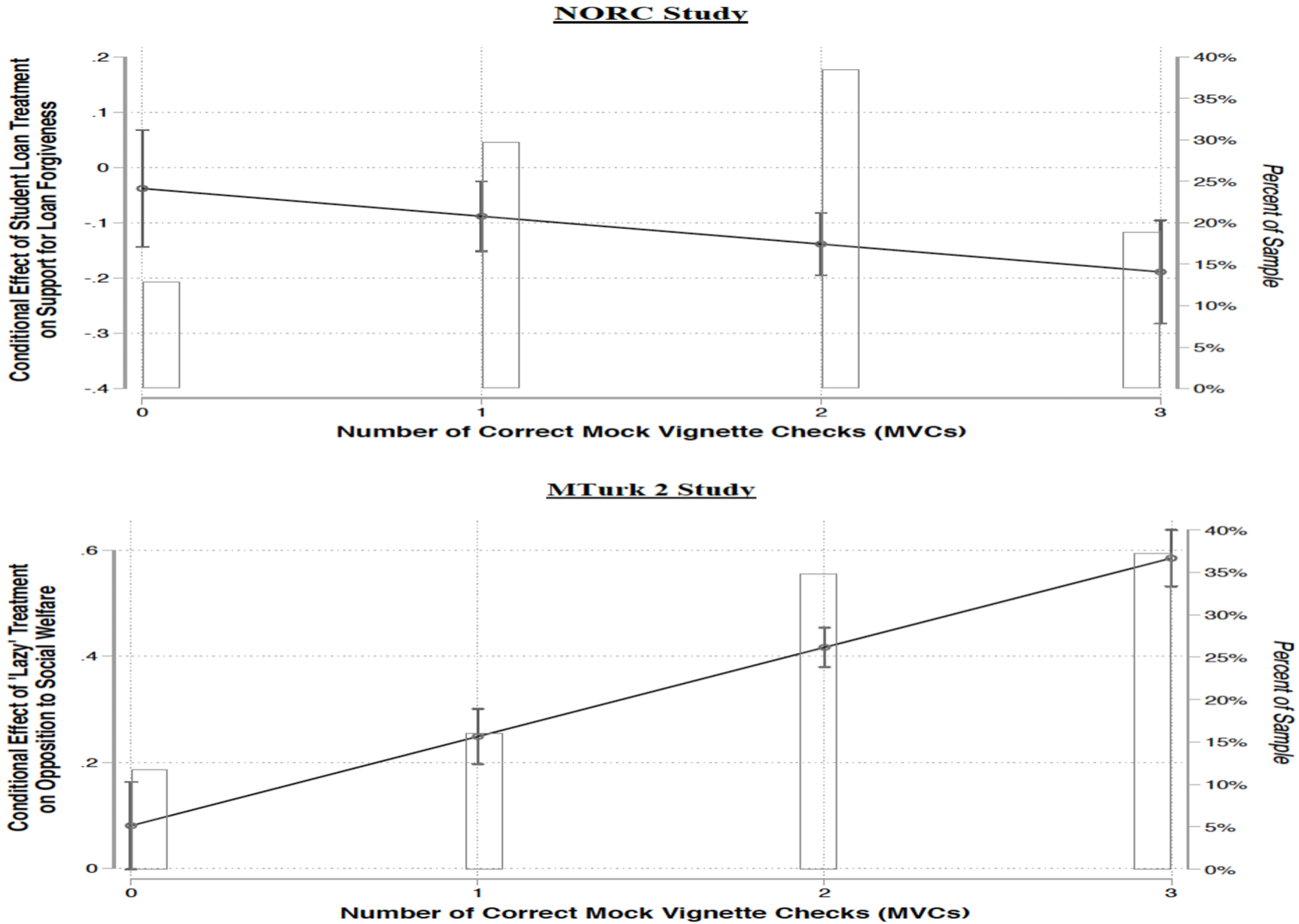
TABLE 1. Overview of Samples, Mock Vignettes, and Experiments

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

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

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





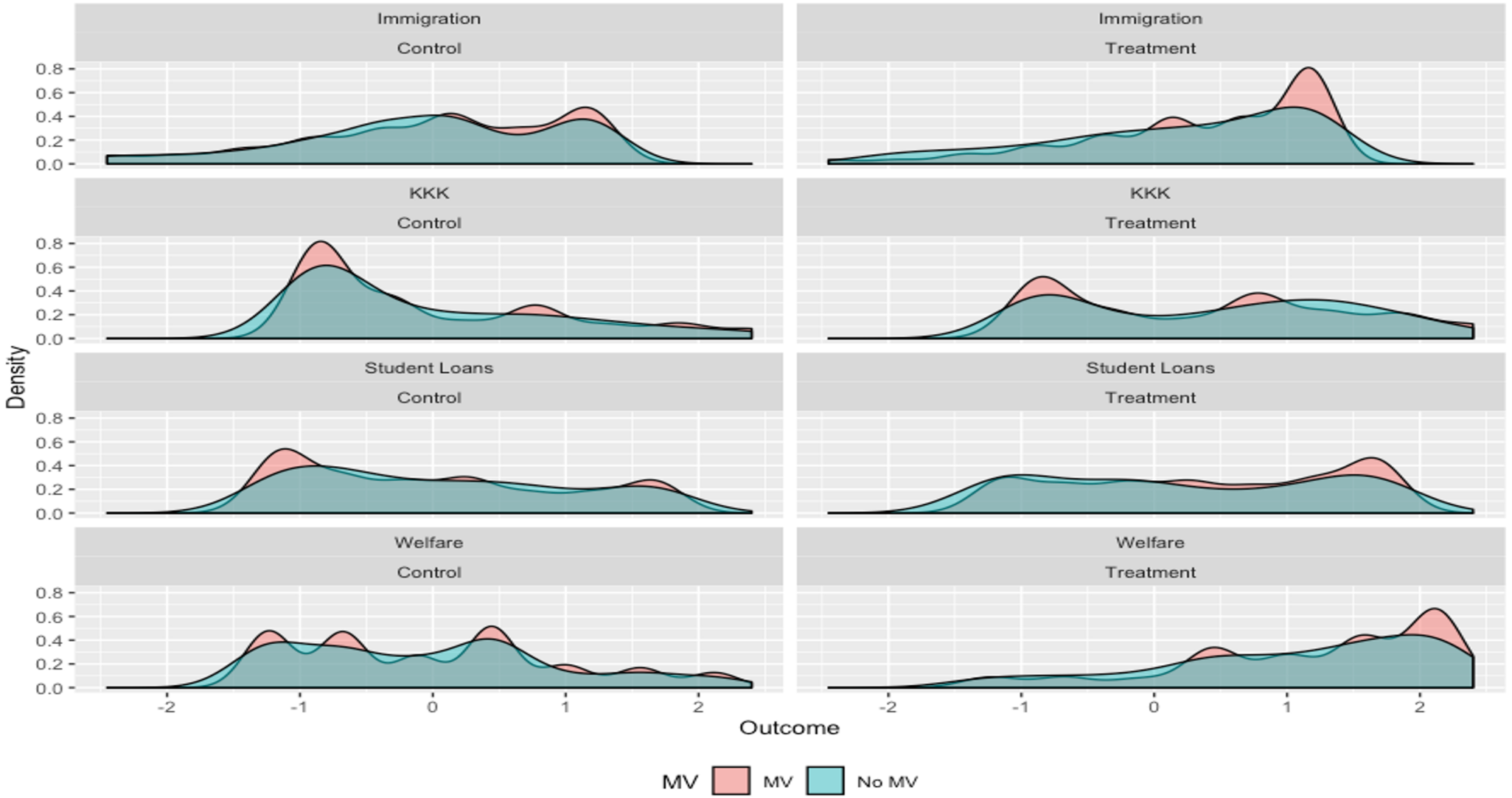
Notes: Figure displays treatment effect estimates for "Student Loan Forgiveness" experiment (top panel) and "Welfare Deservingness" experiment across performance on the mock vignette check scale. 95% confidence intervals shown. Total N=744 (NORC) and 804 (MTurk Study 2).

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).

MVs Do Not Significantly Alter Effects



Demographics of Passers

TABLE D1. Demographic Predictors of MVC Performance

	Mock Vignette Check (Binary)		Mock Vignette Check (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 from 0 to 1. 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 so that higher values indicate more Republican and conservative,

FIGURE I1: Changes in *t* with Better MVC Performance