Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents

Alexander Coppock and Oliver A. McClellan[*]

August 26, 2018

Abstract

Researchers have increasingly turned to online convenience samples as sources of survey responses that are easy and inexpensive to collect. As reliance on these sources has grown, so too have concerns about the use of convenience samples generally and Amazon's Mechanical Turk in particular. We distinguish between "external validity" and theoretical relevance, with the latter being the more important justification for any data collection strategy. We explore an alternative source of online convenience samples, the Lucid Fulcrum Exchange, and assess its suitability for online survey experimental research. Our point of departure is Berinsky et al. (2012), which compares Amazon's Mechanical Turk to national probability samples in terms of respondent characteristics and treatment effect estimates. We replicate these same analyses using a large sample of survey responses on the Lucid platform. Our results indicate that demographic and experimental findings on Lucid track well with national benchmarks, with the exception of experimental treatments that aim to dispel the "death panel" rumor regarding the Affordable Care Act. This exception points to the possibly time-bound nature of some survey experimental effects. We conclude that Lucid can serve as a drop-in replacement for many scholars currently conducting research on Mechanical Turk or other similar platforms.

The use of online convenience samples for experimental research has exploded in recent decades, with far-reaching and mostly positive consequences for scholarship in the social sciences. Due to its low cost and quick turnaround time, Amazon's Mechanical Turk (MTurk) in particular has become a popular testing ground for many social scientific hypotheses. Where once researchers may have only speculated about a causal process, now they can test, refine, and retest in short order.

[*] Alexander Coppock is Assistant Professor of Political Science, Yale University (alexander.coppock@yale.edu). Oliver A. McClellan is a PhD Student in Political Science, Columbia University (oam2112@columbia.edu). The authors received no compensation for this study and the analyses reported here were conducted independently. This study was reviewed and approved by the Institutional Review Board of Columbia University (IRB-AAAQ7500)

That said, convenience samples in general and MTurk in particular may be inappropriate for some experimental research questions. Concerns about convenience samples largely fall into three categories. First, some researchers are concerned about how treatment effect heterogeneity might affect the interpretation of studies that seek to estimate average causal effects (Krupnikov and Levine, 2014). If subjects on MTurk harbor different latent responses to treatment compared with those in the true population of interest, then treatment effect heterogeneity may impede extrapolation from Mechanical Turk to other populations. While recent meta-analyses of studies conducted on both Mechanical Turk and national probability samples (Mullinix et al., 2015; Coppock, 2017; Coppock et al., 2017) have found generally high degrees of correspondence across platforms, successes in some particular domains do not ensure successes in others.

Second, some researchers worry about pronounced demand effects. Behrend et al. (2011) show that MTurk responses are slightly more susceptible to social desirability bias than other samples. Others are concerned that Mechanical Turk respondents perceive a conditional relationship between the answers they give and the pay they earn. Bullock et al. (2015) have shown that the political beliefs (as expressed by a survey response) can be affected by payments for "correct" responses. Rightly or wrongly, subjects on MTurk may believe that they will earn more money if they respond in a particular manner. We note that recent experimental evidence has found little to no evidence of demand effects White et al. (2018); Mummolo and Peterson (2018); De Quidt et al. (2018), even when indicating the investigators' preferred responses with heavy-handed messages.

Finally, some scholars are concerned that MTurk is "overfished" and that many respondents have become professional survey takers (Rand et al., 2014; Chandler et al., 2015). Stewart et al. (2015) estimate the pool of active MTurk respondents for a given lab to be approximately 7,300 subjects at any one time. MTurk subjects have access to websites where they share information about academic surveys, which is particularly troubling for experiments in which subjects' compensation depends on how they respond. MTurk participants share advice on how to maximize these payoffs on sites such as Turkopticon (turkopticon.ucsd.edu) or Turkernation (turkernation.com).

Regardless of whether any or all of these concerns about MTurk hold in a particular research scenario, it behooves social scientists to consider other sources of subjects, if only to hedge bets through diversification. In our view, a source of convenience should satisfy four technical desiderata. The pool of subjects should be large and diverse, respondents should not to be able to coordinate, samples should approximate target populations well (if applicable), and costs should be kept to a minimum. We emphasize that even if a convenience sample achieves these four goals, the overriding requirement for any sample – convenience or otherwise – should be theoretical relevance.

In this paper, we consider when convenience samples should be used for social scientific research and demonstrate the viability of a new source of subjects – Lucid – as an alternative to Mechanical Turk.

1 Convenience samples and theoretical relevance

Before turning to the specifics of the Lucid platform, we consider the conditions under which researchers should turn to online convenience samples as sources of subjects.[1] A major distinction to be drawn is between descriptive studies and experimental studies. Researchers interested in describing features of a population based on a sample should be very cautious when generalizing from convenience samples. While it is commonplace in the popular media to conduct opinion polls using convenience samples of viewers or listeners (Kent et al., 2006), most descriptive work in political science uses explicit random sampling or reweighting techniques to target population quantities (Park et al., 2004). Although even extremely idiosyncratic convenience samples (Gelman et al., 2016, e.g., Xbox users) can sometimes produce estimates that turn out to have been accurate, we would not generally recommend using Lucid (or any convenience sample) when the goal is to describe the features of a national population.

Experimental studies, by contrast, often seek to estimate the sample average treatment effect (SATE), though other estimands (such as SATEs conditional on pretreatment covariates) are also common. Estimates of the SATE are said to exhibit strong internal validity if the standard experimental assumptions are met; this logic extends to samples obtained from Lucid.[2]

The question of whether a particular convenience sample should be used for survey experiments depends, of course, not only on internally validity, but also on whether the SATE is worth estimating at all. The SATE is often contrasted with the population average treatment effect (PATE), and the SATE is said to exhibit poor external validity if the SATE is different from the PATE. We do not share this view of external validity. The PATE and the SATE are different estimands, and estimates of each may be more or less useful depending on the target of inference.[3]

In our view, the choice to use a convenience sample should depend on whether the SATE is *relevant for theory.* Whether a given SATE is relevant for theory will doubtless be a matter of debate in any substantive area. If the goal is to study the effect of a English-language newspaper article on political opinion, the SATE from a convenience sample of French-only monolinguals would not be relevant for theory, for the simple reason that the hypothesized causal process would not take place because the subjects do not speak English. A heuristic for determining whether a SATE is relevant for theory is to consider whether the theory's predictions also apply to that sample, *not* whether that sample is "representative" of some different population. Our guess is that if a theory applies to the national population (i.e., adult Americans), it should usually apply to a subset of that population (i.e., adult Americans on Lucid), though we grant there may be exceptions. As

[1] Sometimes researchers employ people encountered online (through Mechanical Turk or other platforms) as research assistants for coding text or other tasks. In this paper, we only consider the use of convenience samples for survey experiments.

[2] See Gerber and Green (2012, chp. 2) for a discussion of the three core assumptions required for internally valid inference in an experiment.

[3] As a discipline, we often speak of "the" PATE as if there is only one, but of course the "P" in PATE could refer to any well-defined population, such as Bostonians in 1983.

it happens, survey experimental SATE and PATE estimates are frequently quite similar (Mullinix et al., 2015; Coppock, 2017; Coppock et al., 2017), and the main explanation for this finding seems to be low treatment effect heterogeneity in response to the sorts of treatments studied by social scientists in survey experiments. Boas et al. (N.d.) report a similar finding from a comparison of subjects recruited via Facebook, Qualtrics, and Mechanical Turk.

We emphasize that describing a SATE as relevant for theory does not require an assumption that the SATE is equal to the PATE. If a SATE is relevant for theory, then it is interesting in its own right, regardless of whether the SATE and the PATE are the same number (or even have the same sign). Researchers always have to defend the provenance of their samples; defending convenience samples means specifically arguing that the theory under examination applies to the people in the convenience sample.

2 Subject Recruitment

Lucid is the largest marketplace for online "sample" nationwide. Providers direct respondents to Lucid, which then redirects subjects to purchasers, typically market research firms. The providers typically compensate survey takers in cash, gift cards, or reward points. As soon as subjects enter the marketplace (and every subsequent 3 months), their demographic characteristics (age, gender, ethnicity, race, education, income, and ZIP code) are measured using the U.S. Census question wordings and response options. Because Lucid does not store any personally identifying information (beyond these demographics), any such information cannot be passed on to the researcher. Approximately 375,000 unique respondents pass through the exchange each day; in 2015, Lucid managed 30 million unique respondents[4]. Lucid can construct demographically-targeted sets of respondents using a combination of quota sampling and screening questions. For example, Flores and Coppock (2017) obtained 2,866 Spanish-English bilingual subjects on Lucid using a custom screening question that asked subjects to self-identify as bilingual. Approximately 95% of all subjects are recruited using a double opt-in procedure: they opt in to being a panel member and opt in to participating in a specific survey. For a 10-minute survey delivered to a group of subjects quota sampled to match census demographics, researchers can expect to pay approximately $1 per completed response as of 2018. See Graham (2018) for a Lucid sample constructed in this manner.

Just like Mechanical Turk (Mason and Suri, 2012; Paolacci and Chandler, 2014), the composition of the pool of survey respondents on Lucid changes over time as both providers and respondents enter and exit the exchange. As we only have a single sample of 3,504 subjects obtained in March of 2016, we cannot empirically assess the extent of overtime variation. However, our concerns about temporal differences in sample composition are assuaged somewhat by the ability to quota sample.

[4] These figures obtained via private correspondence. Lucid tracks unique respondents through a combination of IP address and provider-maintained unique identifiers. While this process is not perfect, Lucid attempts to deduplicate using a set of geographic and demographic checks.

Quota sampling ensures that the marginal (but not necessarily joint) distributions of demographic characteristics match predetermined targets. What remains are the (possibly unobservable) non-demographic characteristics that may drift over time. Such drift would pose a challenge for Lucid if these characteristics interact with treatment in important ways. We consider overtime drift as one explanation for the nonreplication of one of the five studies we replicate below.

Because much of the concern over the use of MTurk has been the professionalization of subjects on the platform, we attempted to assess the survey-taking behavior of subjects on Lucid, as shown in Table 1. Respondents report taking an average of 4.28 surveys per month. However, 98 percent of respondents report taking fewer than one survey per day; the average number of surveys per month among these respondents is 2.43. The vast majority of subjects (94%) take surveys at home, and the majority are compensated directly in dollars or in some form of points program. We asked our subjects to report the dollar value of their expected compensation, but we suspect that some subjects entered the number of points they expected to receive. Unconditionally, the average compensation amount that subjects reported expecting was $5.01, but if we trim off responses that are implausible (greater than $20.00), we obtain the more reasonable figure of $1.16.

Table 1: Lucid sample survey behavior

	Lucid
Number of surveys taken in the last month	4.28 (0.40)
All Responses	4.28 (0.40)
Responses <= 30	2.43 (0.07)
Survey Location	
Home (%)	93.51 (0.42)
Work (%)	3.32 (0.30)
Public place such as a library (%)	1.34 (0.19)
Other (%)	1.83 (0.23)
Survey Compensation Type	
US dollars (%)	55.00 (0.84)
Website points (%)	36.19 (0.81)
Bitcoin (%)	1.23 (0.19)
Other national currency (%)	1.23 (0.19)
No compensation (%)	6.35 (0.41)
Compensation amount	
All Responses	$5.01 (0.30)
Responses <= $20	$1.16 (0.04)
N	3,504

Standard errors in parentheses where applicable.
All entries are self-reported figures.

3 Baseline Characteristics

Before comparing the SATEs obtained on Lucid to those obtained on MTurk and on probability samples, we assess the distribution of baseline characteristics like demographics, political attitudes, and psychological traits. Figure 1 presents standardized demographic means on Lucid, MTurk, and the ANES 2012, where we standardize by the ANES 2012 mean and standard deviation.[5]

Demographics

In terms of gender, education, age, and income, the Lucid sample comes closer to the ANES 2012 benchmarks than does the MTurk sample. The Lucid sample was 52% female, much closer to the Census value of 50.8% than the 60% female sample collect on MTurk. The mean number of years of education on Lucid (14.2) is higher than the approximately 13.5 years recorded by the ANES survey, but is closer than MTurk sample estimates. Both mean and median incomes are lower on Lucid than among the face-to-face sample, but are higher than in the MTurk sample. Both of the internet samples overrepresent whites relative to nonwhites, but this distortion is smaller on Lucid. The regional balance on Lucid comes very close to the 2012 ANES, whereas the MTurk sample appears to overrepresent southerners.

Out of the 11 demographic variables that are measured for both Lucid and MTurk, the Lucid mean is closer to the ANES mean in 9 instances, 5 of which are statistically significant.[6]

Politics

Voter registration and turnout seem to vary somewhat across samples, with the Lucid sample corresponding more closely to the 2012 ANES baseline for voter registration and voter turnout.[7] Political party affiliation seems to track closely across samples, though the Lucid mean of 3.7 is identical to that collected in the 2012 ANES, while the MTurk average is slightly lower at 3.5. We see important variation with regard to respondents' ideologies: respondents on MTurk are markedly more liberal than respondents found on Lucid or the ANES.

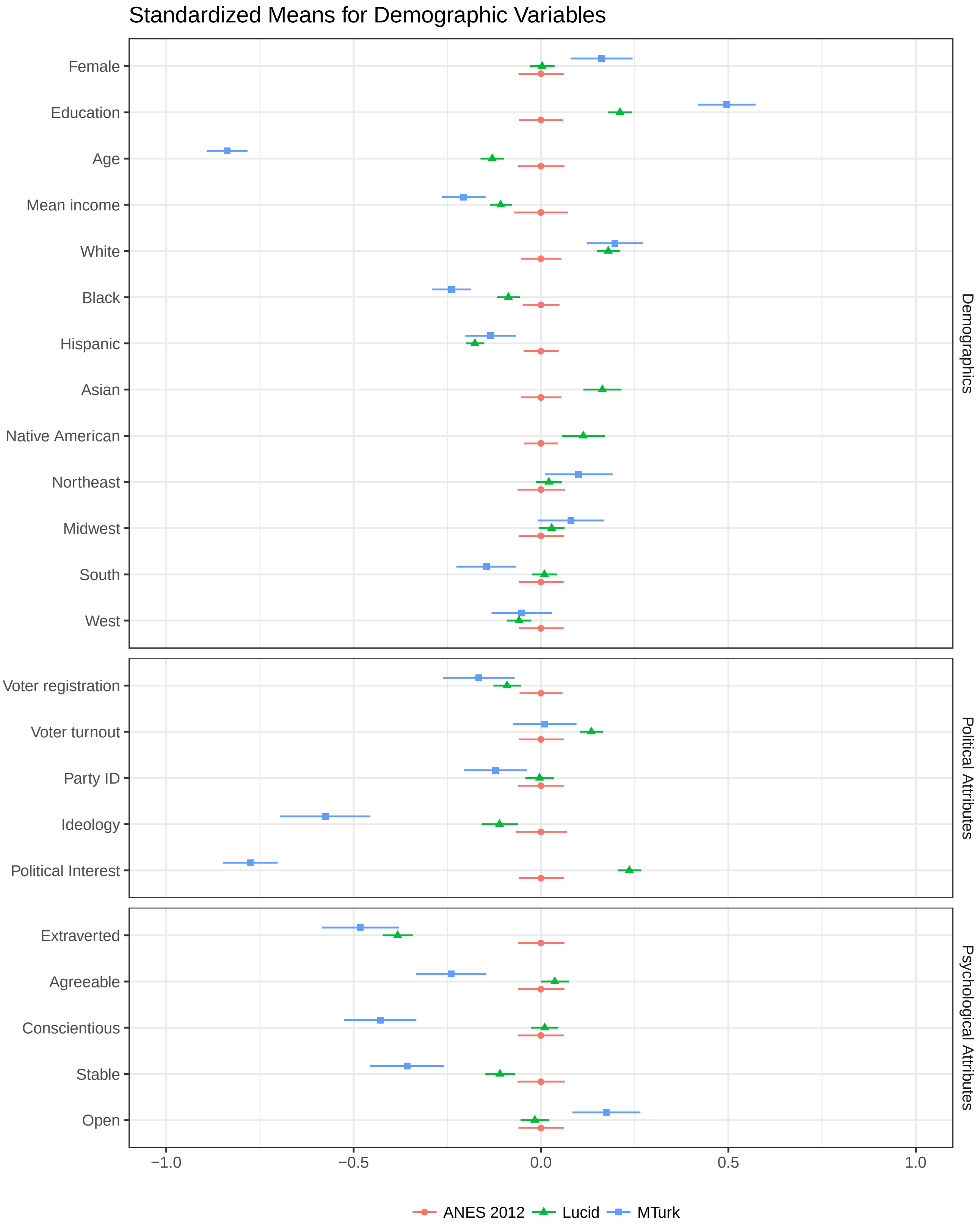
Interest in politics varies across samples. On average, MTurk respondents have the least interest in politics, while Lucid respondents have the most. The difference between Lucid and MTurk is large, about 1.2 points on the 5-point political interest scale. This trend is reversed for political knowledge, included in Appendix A, Table 2. MTurk respondents scored higher on political knowl-

[5] Corresponding tables that include comparisons to additional samples are shown in Appendix A. For the 2008 ANES, we use the post-election, post-stratified weight to analyze the data, which takes into account attrition between the pre- and post-election waves. For the 2012 ANES, we use the post-stratified, face-to-face weight, since we analyze only the face-to-face sample. For the ANES panel data, we use the post-stratified weight from Wave 11, the latest wave from which we analyze data.

[6] See the online appendix for the boostrapping procedure we used to conduct these tests.

[7] All surveys here examined significantly overstate both turnout and registration compared to official government statistics. For a detailed discussion of this phenomenon, see DeBell et al. (2018).

Figure 1: Standardized Means for Demographic Variables



edge then did respondents on the ANES Panel while Lucid respondents scored nearly identically. We speculate that MTurk's strong performance across our political interest and knowledge questions may be due to MTurk respondents being familiar with the knowledge batteries employed in many political science studies conducted on MTurk. Lucid is significantly closer to the ANES 2012 on party identification, ideology, and political interest, while MTurk is significantly closer for voter turnout.

The average policy preferences held by each of the samples in a variety of domains are also shown in Appendix A, Table 2. These estimates are generally consistent across samples, with Lucid polling slightly more conservatively than MTurk. This fits with the ideological differences we observe between the two samples. MTurk respondents are the least likely to favor prescription drug benefits for seniors, possibly because MTurk respondents are younger on average.

Psychology

Finally, we compare Lucid, MTurk, and the ANES in terms of the "Big 5" personality indices, as measured by the Ten Item Personality Inventory (Gosling et al., 2003), which has been shown to correlate with a host of other characteristics including political views (Gerber et al., 2010). The Lucid sample tracks very well with the CCES, CCAP, and ANES 2012 on all five personality traits, perhaps slightly outperforming the MTurk sample on Conscientiousness and Stability. This correspondence is encouraging, as nothing about the quota sampling process used by Lucid should guarantee similarity to national samples on psychological traits. Formal hypotheses tests demonstrate that Lucid is significantly closer to the ANES 2012 than MTurk on all five traits.

4 Experiments

Thus far, we have compared the performance of samples collected on Lucid and MTurk with respect to baseline levels of their demographic, political, and psychological profiles. However, our main concern is the performance of Lucid for survey experiments. We replicated five separate survey experiments originally conducted on other platforms on our Lucid sample. In our reading, the causal theories elaborated in the original studies should travel to the Lucid sample as well as they do to the U.S. national populations or to the MTurk population, so we will conclude that Lucid does well if we obtain qualitatively similar average effect estimates.

For space reasons, we provide brief descriptions of each experiment in the main text along with summary figure comparing the estimated treatment effects across sample. In the Welfare, Asian Disease, Kam and Simas, Hiscox and Berinsky facets found in Figure 2, we present standardized treatment effect estimates, where we have scaled the outcome variables for Lucid and MTurk by the mean and standard deviation of the original experiments. The Berinsky facet does not include an MTurk estimate since it has not been previously replicated on an MTurk sample. Fuller descriptions of our procedures and results (including treatment and outcome question wordings as well as regression tables of our results) are available in the online appendix. We did not pre-register our analyses because we follow the analysis strategies of the original authors. Again following the original authors, we drop subjects with missing or don't know outcomes.[8] In all cases, we estimate HC2 robust standard errors to construct 95% confidence intervals and conduct hypothesis tests.

Experiment 1: Welfare Spending

Our first experiment replicates a classic question wording experiment. Control subjects are asked whether we are spending too little, about right, or too much on "welfare." Treatment subjects are asked the same question about "Assistance to the poor" or "Caring for the poor." The General Social Survey (GSS) has conducted this experiment every other year since 1984; we use the 2014 GSS estimate as the baseline result. This experiment behaves on Lucid much as it does on MTurk and the GSS – a large increase in support for redistribution when the question is phrased as assistance or caring for the poor rather than as "welfare."

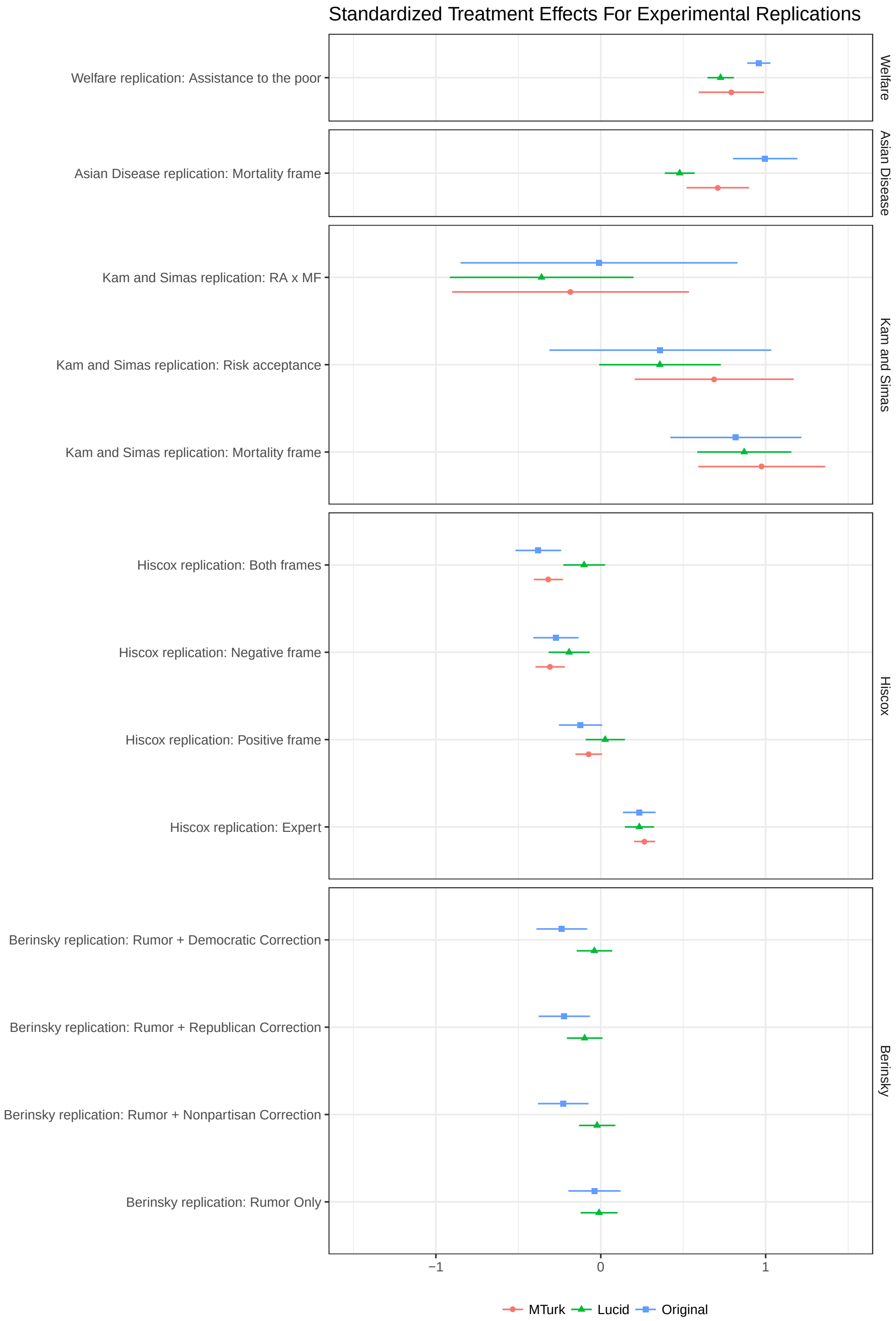
Experiment 2: Asian Disease Problem

Our second experiment is also a classic, this time of the behavioral economics literature. Tversky and Kahneman (1981) show that people take the riskier option when in a "loss frame" rather than a "gain frame." Subjects are asked to "Imagine that your country is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed." Subjects in the the control condition are told "If Program A is adopted, 200 people will be saved. If Program B is adopted, there is one-third probability that 600 people will be saved, and two-third probability that no people will be saved." Subjects in the treatment group (the "mortality frame") are told: "If Program A is adopted, 400 people will die. If Program B is adopted there is one-third probability that nobody will die, and two-third probability that 600 people will die."

Across all three samples (the original experiment was conducted in a classroom setting among undergraduates), the treatment has average effects in the same direction, with subjects in the mortality (loss) frame far more likely to choose the probabilistic (risky) outcome, though the magnitudes of the effects do differ substantially by sample. Lacking a national sample benchmark, it is unclear how to grade Lucid's performance relative to MTurk, though we would argue that the qualitative conclusions drawn from the experiment are the same across all samples.

[8] We do not assume that missingness is random, rather, we make the assumption that treatments do not cause missingness. That is, we assume all subjects are either "Always-Reporters" or "Never-Reporters," and therefore technically speaking, our estimand is the SATE conditional on reporting. Regressions predicting missingness from treatment assignment are all nonsignificant, bolstering (but not proving) the Always-Reporters assumption (Gerber and Green, 2012, chp. 7).

Figure 2: Summary of Experimental Comparisons Across Samples



Experiment 3: Framing and Risk

Our third experiment replicates Kam and Simas (2010), who show that risk acceptance correlates with choosing the risky option in an Asian Disease-type experiment, but that the treatment effect of the mortality frame does not vary appreciably with risk acceptance.[9] This finding was replicated in both MTurk and Lucid samples. Receiving the mortality frame increases the likelihood of selecting the probabilistic choice. Risk acceptance correlates with choosing the risky option, but does not moderate the effect of treatment. Both MTurk and Lucid samples are able to replicate estimates of (the lack of) heterogeneous treatment effects.

Experiment 4: Free Trade

Study 4 is a replication of Hiscox (2006), which measured the effects of positive, negative, and expert opinion frames on support for free trade. The study employed a 2 x 4 design. The first factor is the Expert treatment, which informed subjects that economists are nearly unanimously in favor of free trade. The second factor is the valence frame, which highlights positive, negative, or both positive and negative impacts of free trade on the economy and jobs. Control subjects saw no frames before proceeding to the outcome question answered by all subjects: “Do you favor or oppose increasing trade with other nations?” The Expert frame increases support for free trade in all examined samples, while the positive frame has negligible (or even negative) effects and the negative frame has unambiguously negative effects. In both the original sample and the MTurk sample, the combination of the positive and negative frames decreased support. Overall, the studies yield similar experimental estimates.

Experiment 5: Health Care Rumors

We conclude our set of five experiments with a note of caution. We attempted to replicate Berinsky’s 2017 experiment on belief in rumors surrounding the Affordable Care Act, specifically the false rumor that the ACA would create “death panels” that would make end-of-life decisions for patients without their consent. In the original experiment (conducted in 2010 on a sample provided by Survey Sampling International, or SSI), a large portion of the sample believed the rumor, and corrections delivered by Republicans, Democrats, and Nonpartisan groups all were effective in correcting false beliefs.

When we replicated the experiment on Lucid, we found a similar level of baseline belief in the rumor. On a -1 to 1 scale (with 0 indicating the respondent was “not sure”), average levels of

[9] Berinsky et al. (2012) analyze the original and their replication using a probit model but we use ordinary least squares (OLS). While some analysts prefer to use nonlinear models like logit or probit when the dependent variable is binary, in an experiment, OLS (without adjustment) is unbiased for the average treatment effect (or, as in this case, the conditional average treatment effects). See Gerber and Green (2012, chp. 2) for a textbook proof. The substantive conclusions do not change in any way if we use probit regression.

belief were -0.17 on Lucid, compared with -.19 in the original. However, none of the corrections (with the possible exception of the Republican correction) appear to have had effects as large as was documented in the original. It could be that the Lucid sample is uniquely impervious to these corrections, but that explanation is hard to reconcile with the fact that the original sample was an online convenience sample much like Lucid. We think that a more plausible explanation for this divergence is that opinion on the ACA has hardened in the six intervening years between the original implementation and when we conducted our replication. These results underline that treatment effects can both vary across individuals within the same time period and across time periods within individuals.

5 Discussion

The surge in research conducted online has many positive benefits. Researchers can pilot quickly and make adjustments to strengthen their designs. Because online convenience samples are inexpensive to collect, researchers can more easily conduct experiments at scale. Online surveys have also lowered the barriers to entry for early-career scholars. The dramatic increase in the use of online convenience samples raises at least two questions. First, for which research tasks are online convenience samples appropriate? Second, when convenience samples are appropriate, is Mechanical Turk the best option, or are there alternatives?

Regarding the question of when convenience samples are appropriate, we have offered the point of view that a sample must be "relevant for theory." Theoretical relevance is a separate construct from external validity. Loosely, external validity is the extent to which we would obtain similar answers if we did the same experiment on new samples, possibly drawn at random from a well-defined population. Theoretical relevance, by contrast, is about whether the theory's predictions also hold for the convenience sample – whether or not the sample average treatment effect exactly equals the population average treatment effect. We emphatically do not draw this distinction in order to imply that any experiment conducted on a convenience sample is relevant for theory.

In our five experiments, Lucid performed remarkably well in recovering estimates that some close to the original estimates. In most cases, our estimates matched the original in terms of sign and significance. In zero cases did we recover an estimate that was statistically significant and had the opposite sign from the original. We think that the best explanation for this pattern is low treatment effect heterogeneity, which is another way of saying that the causal theories laid out in the original papers extend in a straightforward way to the Lucid sample.

Among our five experiments, we have one instance of a replication "failure." In no way do we think our results contradict or overturn those reported by Berinsky (2017). Instead, we suspect that the correction no longer works because times have changed since the original experiment. While this line of reasoning is admittedly post-hoc, one might argue that the Lucid sample was not relevant for the rumor correction theory because by 2016, attitudes and opinions about Barack

Obama were strongly held by most Americans. If so, this heterogeneity in response to treatment is a feature of Americans generally and not a unique feature of the special subset of Americans who take surveys on Lucid. Of course, we can neither confirm nor disconfirm this speculation in the absence of a massive replication on a national sample.

Regarding the second question of how to choose among sources of convenience samples, we believe we have shown that subjects obtained via Lucid can serve as a drop-in replacement for subjects recruited on Mechanical Turk. Lucid boasts a much larger pool of subjects than Mechanical Turk; the risk of cooperation among subjects is minimal given their diverse sources; subjects are less professionalized; subjects are more similar to national benchmarks in terms of their demographic, political, and psychological profiles. Experimental results obtained on Lucid are solidly in line with the results obtained on other platforms.

Some notes of caution. First, the MTurk responses summarized by Berinsky et al. (2012) were obtained in 2010; it is possible that MTurk has changed and if these same studies were reconducted on MTurk now, the estimates could be closer (or less close) to the probability sample benchmarks. For example, Berinsky et al. (2012)'s MTurk sample is 60% women; contemporary MTurk samples are closer to 48% women. Second, researchers have developed tools to implement a wide variety of studies on MTurk. For example, the `MTurkR` software (Leeper, 2015) makes it easy to implement panel studies on MTurk. Similar tools have not been developed for Lucid, so some researchers would face significant costs of changing their workflows.

Lastly, we note that Mechanical Turk survey respondents are among the very-best studied human beings on the planet. While we advocate in this paper that scholars seek out new sources of survey respondents, we recognize that the knowledge we have about MTurk workers is valuable. As a research community, we have honed our understanding about how these people respond to incentives, question wordings, and experimental stimuli. We know how they respond to attention checks and distraction tasks. Journal editors and peer reviewers are already familiar with the strengths and weaknesses of MTurk data. Diversifying our subject pools will necessarily involve learning how other online samples are similar and different. While we are reassured that on most dimensions, Lucid data appear to equal or outperform Mechanical Turk data, we also recognize that changing data sources does not come without costs.

References

Behrend, Tara S, David J Sharek, Adam W Meade and Eric N Wiebe. 2011. "The Viability of Crowdsourcing for Survey Research." *Behavior Research Methods* 43(3):800–813.

Berinsky, Adam J. 2017. "Rumors and Health Care Reform: Experiments in Political Misinformation." *British Journal of Political Science* 47(2):241–262.

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Boas, Taylor C., Dino P. Christenson and David M. Glick. N.d. "Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics." *Political Science Research and Methods*. Forthcoming.

Bullock, John G., Alan S. Gerber, Seth J. Hill and Gregory A. Huber. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10(4):519–578.

Chandler, Jesse, Gabriele Paolacci, Eyal Peer, Pam Mueller and Kate A. Ratliff. 2015. "Using Nonnaive Participants Can Reduce Effect Sizes." *Psychological Science* 26(7):1131–1139.

Coppock, Alexander. 2017. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* . Forthcoming.

Coppock, Alexander, Thomas J. Leeper and Kevin J. Mullinix. 2017. "The Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." *Unpublished manuscript* .

De Quidt, Jonathan, Johannes Haushofer and Christopher Roth. 2018. "Measuring and Bounding Experimenter Demand." *American Economic Review* . Forthcoming.

DeBell, Matthew, Jon A Krosnick, Katie Gera, David S Yeager and Michael P McDonald. 2018. "The Turnout Gap in Surveys: Explanations and Solutions." *Sociological Methods & Research* p. 0049124118769085.

Flores, Alejandro and Alexander Coppock. 2017. "Do Bilinguals Respond More Favorably to Candidate Advertisements in English or in Spanish?" *Unpublished Manuscript* .

Gelman, Andrew, Sharad Goel, Douglas Rivers, David Rothschild et al. 2016. "The Mythical Swing Voter." *Quarterly Journal of Political Science* 11(1):103–130.

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

Gerber, Alan S., Gregory A. Huber, David Doherty, Conor M. Dowling and Shang E. Ha. 2010. "Personality and Political Attitudes: Relationships Across Issue Domains and Political Contexts." *American Political Science Review* 104(01):111–133.

Gosling, S.D., P. J. Rentfrow and W. B. Jr. Swann. 2003. "A Very Brief Measure of the Big-Five Personality Domains." *Journal of Research in Personality* 37:504–528.

Graham, Matthew H. 2018. "Self-Awareness of Political Knowledge." *Political Behavior* . Forthcoming.

Hiscox, Michael J. 2006. "Through a Glass and Darkly: Attitudes Toward International Trade and the Curious Effects of Issue Framing." *International Organization* 60(03):755–780.

Kam, Cindy D. and Elizabeth N. Simas. 2010. "Risk Orientations and Policy Frames." *The Journal of Politics* 72(2):381–396.

Kent, Michael L., Tyler R. Harrison and Maureen Taylor. 2006. "A Critique of Internet Polls as Symbolic Representation and Pseudo-Events." *Communication Studies* 57(3):299–315.
URL: *https://doi.org/10.1080/10510970600845941*

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

Leeper, Thomas J. 2015. *MTurkR: Access to Amazon Mechanical Turk Requester API via R.* R package version 0.6.5.1.

Mason, Winter and Siddharth Suri. 2012. "Conducting behavioral research on Amazon's Mechanical Turk." *Behavior research methods* 44(1):1–23.

Miratrix, Luke W, Jasjeet S Sekhon, Alexander G Theodoridis and Luis F Campos. 2018. "Worth Weighting? How to Think About and Use Weights in Survey Experiments." *Political Analysis* pp. 1–17.

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

Mummolo, Jonathan and Erik Peterson. 2018. "Demand Effects in Survey Experiments: An Empirical Assessment." *Unpublished Manuscript* .
URL: *https://scholar.princeton.edu/jmummolo/publications/demand-effects-survey-experiments-empirical-assessment*

Paolacci, Gabriele and Jesse Chandler. 2014. "Inside the Turk: Understanding Mechanical Turk as a participant pool." *Current Directions in Psychological Science* 23(3):184–188.

Park, David K., Andrew Gelman and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-level Estimates from National Polls." *Political Analysis* 12(4):375–385.

Rand, David G, Alexander Peysakhovich, Gordon T Kraft-Todd, George E Newman, Owen Wurzbacher, Martin A Nowak and Joshua D Greene. 2014. "Social Heuristics Shape Intuitive Cooperation." *Nature Communications* 5.

Stewart, Neil, Christoph Ungemach, Adam JL Harris, Daniel M Bartels, Ben R Newell, Gabriele Paolacci and Jesse Chandler. 2015. "The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers." *Judgment and Decision Making* 10(5):479.

Tversky, Amos and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211(4481):453–458.

White, Ariel, Anton Strezhnev, Christopher Lucas, Dominika Kruszewska and Connor Huff. 2018. "Investigator Characteristics and Respondent Behavior in Online Surveys." *Journal of Experimental Political Science* 5(1):56–67.

Appendix A

Where survey weights were provided, all entries in the Appendix tables represent weighted means. For the ANES Panel study, we use the post-stratified Wave 11 weights, as Wave 11 is the latest wave from which we analyze data. The Lucid sample is unweighted. While we could, in principle, reweight the Lucid data to Census targets, we follow what appears to be standard practice with convenience samples and do not weight our estimates (Miratrix et al., 2018). The results for the ANES 2008 and ANES Panel in some cases differ from those presented in Berinsky et al. (2012), as corrected datasets for these two surveys have been released since the publication of Berinsky et al. (2012).

Table 1: Comparing sample demographics

	Internet samples			*Face-to-face samples*		
	Lucid	*MTurk*	*ANESP*	*CPS 2008*	*ANES 2008*	*ANES 2012*
Female (%)	52.15 (0.85)	60.07 (2.09)	51.80 (2.19)	51.67 (0.18)	55.29 (1.37)	52.01 (1.53)
Education (mean years)	14.16 (0.04)	14.88 (0.10)	15.15 (0.12)	13.21 (0.01)	13.50 (0.07)	13.63 (0.07)
Age (mean years)	44.92 (0.28)	32.26 (0.49)	46.39 (0.82)	46.02 (0.06)	46.53 (0.50)	47.25 (0.57)
Mean income	$60,896.33 ($833.41)	$55,331.82 ($1,659.29)	$66,710.48 ($1,846.14)	$62,255.92 ($149.25)	$66,158.53 ($1,437.46)	$66,948.92 ($2,039.72)
Median income	$47,500	$45,000	$55,000	$55,000	$55,000	$52,500
Race						
White (%)	78.87 (0.69)	79.67 (1.72)	78.43 (2.02)	68.50 (0.17)	75.37 (0.97)	70.70 (1.24)
Black (%)	9.11 (0.49)	4.17 (0.85)	10.09 (1.54)	11.67 (0.12)	12.34 (0.63)	11.93 (0.81)
Hispanic (%)	5.41 (0.38)	6.72 (1.07)	5.93 (1.20)	13.66 (0.13)	8.15 (0.47)	10.92 (0.73)
Asian (%)	3.44 (0.31)		2.70 (0.81)	5.02 (0.08)	2.79 (0.51)	1.47 (0.33)
Native American (%)	1.03 (0.17)		1.49 (0.58)	1.07 (0.04)	1.35 (0.31)	0.36 (0.14)
Region of the United States						
Northeast (%)	18.92 (0.66)	21.96 (1.77)	18.76 (1.79)	18.42 (0.14)	14.09 (1.03)	18.10 (1.23)
Midwest (%)	23.79 (0.72)	25.95 (1.87)	23.91 (1.69)	21.91 (0.14)	20.97 (1.17)	22.60 (1.27)
South (%)	37.63 (0.82)	30.13 (1.96)	36.37 (2.15)	36.54 (0.18)	43.50 (1.35)	37.20 (1.46)
West (%)	19.66 (0.67)	19.96 (1.70)	20.96 (1.77)	23.13 (0.15)	21.44 (1.04)	22.10 (1.26)
N	3,504	551	1,058	100,008[1]	2,322	2,054

[1] Income figures derived from CPS Income Supplement, N = 150,799.
Entries are unweighted for Lucid and MTurk and are weighted for ANESP, CPS 2008, ANES 2008 and ANES 2012.
Standard errors in parentheses where applicable.

Table 2: Comparing sample political behavior, traits and opinions

	Internet samples			*Face-to-face samples*		
	Lucid	*MTurk*	*ANESP*	*CPS 2008*	*ANES 2008*	*ANES 2012*
Behavior						
Registered (%)	81.49 (0.66)	78.77 (1.74)	86.66 (1.86)	71.00 (0.17)	86.01 (0.95)	84.75 (1.05)
Voter turnout (%)	76.33 (0.72)	70.64 (1.95)	83.96 (1.83)	73.80 (0.18)[1]	77.44 (1.13)	75.63 (1.31)
Traits						
Party identification (mean on 7-point scale, 7= Strong Republican)	3.73 (0.04)	3.49 (0.09)	3.96 (0.09)		3.73 (0.06)	3.73 (0.06)
Ideology (mean on 7-point scale, 7= Strong conservative)	4.09 (0.04)	3.39 (0.09)	4.45 (0.07)		4.24 (0.05)	4.25 (0.05)
Political Interest[2](mean on 5-point scale, 5 = Extremely interested)	3.62 (0.02)	2.43 (0.04)	3.20 (0.05)		2.94 (0.04)	3.34 (0.04)
Political knowledge	58.39 (0.46)	70.51 (1.03)	59.91 (1.22)			
Opinions						
Prescription drug benefits for seniors (% favor)	74.19 (0.74)	63.52 (2.05)	77.13 (1.73)		79.81 (1.60)	
Universal Healthcare (% favor)	49.69 (0.84)	47.73 (2.13)	43.38 (2.20)		51.07 (1.95)	
Citizenship process for illegal immigrants (% favor)	35.99 (0.81)	38.11 (2.07)	41.39 (2.16)		49.11 (1.95)	
Ban gay marriage (% favor)	28.09 (0.76)	15.61 (1.55)	32.95 (2.03)			
Increase taxes on the rich(% favor)	63.67 (0.81)	61.16 (2.08)	55.51 (2.18)			
Increase taxes on the poor (% favor)	11.42 (0.54)	6.17 (1.03)	6.40 (1.13)			
N	3,504	551	1,058	100,008	2,322	2,054

[1] The Census Bureau incorrectly classified cases of nonresponse as cases of nonvoting. We have corrected this miscoding using listwise deletion, following the methodology presented in DeBell et al. (2018).

[2] ANES 2012 uses alternate wording for political interest; see appendix for exact wording.

Entries are unweighted for Lucid and MTurk, and are weighted for ANESP, CPS 2008, ANES 2008 and ANES 2012.

Standard errors in parentheses where applicable.

Table 3: Comparing sample psychological profiles

	Internet samples					*Face-to-face samples*		
	Lucid	*MTurk*	*CCES*	*CCAP*	*ANESP*	*ANES 2008*	*ANES 2012*	*Kam and Simas (2010)*
Big 5 Personality Index[1]								
Agreeable	0.69 (0.00)	0.64 (0.01)	0.69 (0.01)	0.71 (0.00)			0.69 (0.01)	
Conscientious	0.77 (0.00)	0.69 (0.01)	0.77 (0.01)	0.76 (0.00)			0.77 (0.01)	
Stable	0.64 (0.00)	0.58 (0.01)	0.66 (0.01)	0.67 (0.00)			0.66 (0.01)	
Extraverted	0.49 (0.00)	0.47 (0.01)	0.52 (0.01)	0.52 (0.00)			0.57 (0.01)	
Open	0.67 (0.00)	0.71 (0.01)	0.68 (0.01)	0.70 (0.00)			0.68 (0.01)	
Risk acceptance	0.49 (0.00)	0.51 (0.01)						0.45 (0.01)
Need for cognition	0.57 (0.01)	0.63 (0.01)			0.60 (0.01)	0.56 (0.01)		
Need to evaluate	0.58 (0.00)	0.63 (0.01)			0.56 (0.01)	0.56 (0.01)		
N	3,504	551[2]	1,500	20,000	1,058	2,322	2,054	760

[1] ANES 2012 uses alternate wording for Big 5 Personality Index; see appendix for exact wording.

[2] NFC, NTE and Risk acceptance figures are from MTurk Kam and Simas replication dataset, N = 763.

Entries are unweighted for Lucid, MTurk, and Kam and Simas (2010) and are weighted for ANESP, CPS 2008, ANES 2008, ANES 2012, CCES and CCAP.

Standard errors in parentheses where applicable.

Table 4: Welfare Replications

	Support for Welfare/Assistance			
	Lucid	MTurk	GSS 1984	GSS 2014
	(1)	(2)	(3)	(4)
Treatment: ‘assistance’	0.613***	0.668***	0.693***	0.808***
	(0.034)	(0.085)	(0.048)	(0.030)
Treatment: ‘caring’	0.738***	0.781***		
	(0.033)	(0.080)		
Constant (Control)	1.768	1.695	1.837	1.712
	(0.025)	(0.056)	(0.037)	(0.022)
N	3,294	494	943	2,457
R^2	0.151	0.178	0.179	0.229

*p < .1; **p < .05; ***p < .01

Robust standard errors are in parentheses.

Table 5: Asian Disease Replications

	Preference for the Probabilistic Outcome		
	Lucid	MTurk	Original
	(1)	(2)	(3)
Mortality Frame	0.239***	0.355***	0.498***
	(0.023)	(0.048)	(0.050)
Intercept	0.397	0.260	0.283
	(0.016)	(0.031)	(0.037)
N	1,813	379	307
R^2	0.057	0.128	0.249

*p < .1; **p < .05; ***p < .01

Robust standard errors are in parentheses.

Table 6: Kam and Simas (2010) Replication

	Preference for the Probabilistic Outcome								
	Lucid			MTurk			Original		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mortality Frame	0.346***	0.342***	0.433***	0.439***	0.439***	0.486***	0.405***	0.411***	0.407***
	(0.023)	(0.023)	(0.072)	(0.032)	(0.032)	(0.098)	(0.034)	(0.034)	(0.101)
Risk Acceptance	0.090	0.182**	0.178*	0.299***	0.307***	0.342***	0.176*	0.203*	0.179
	(0.071)	(0.073)	(0.093)	(0.091)	(0.095)	(0.122)	(0.102)	(0.109)	(0.170)
RA X MF			−0.179			−0.092			−0.006
			(0.141)			(0.182)			(0.213)
Intercept	0.261	0.118	0.218	0.104	0.081	0.081	0.240	0.203	0.238
	(0.038)	(0.061)	(0.048)	(0.049)	(0.090)	(0.063)	(0.051)	(0.092)	(0.079)
Covariates	No	Yes	No	No	Yes	No	No	Yes	No
N	1,629	1,629	1,629	766	766	766	752	752	752
R^2	0.120	0.133	0.121	0.204	0.205	0.204	0.166	0.172	0.166

*p < .1; **p < .05; ***p < .01
Robust standard errors are in parentheses.

Table 7: Hiscox Replications

	Support for Free Trade			
	Lucid	MTurk	GfK	Original
	(1)	(2)	(3)	(4)
Expert	0.111***	0.126***	0.087***	0.111***
	(0.021)	(0.015)	(0.020)	(0.024)
Positive Frame	0.013	−0.035*	−0.019	−0.059*
	(0.029)	(0.019)	(0.027)	(0.032)
Negative Frame	−0.092***	−0.146***	−0.170***	−0.130***
	(0.030)	(0.021)	(0.028)	(0.033)
Both Frames	−0.048	−0.151***	−0.109***	−0.181***
	(0.030)	(0.021)	(0.028)	(0.033)
Constant (Control)	0.686	0.784	0.723	0.702
	(0.023)	(0.016)	(0.021)	(0.024)
N	1,811	2,972	2,084	1,578
R^2	0.023	0.046	0.032	0.033

*p < .1; **p < .05; ***p < .01
Robust standard errors are in parentheses.

Table 8: Berinsky (2016) Replication

	Death Panel Rumor Belief (-1 to 1)	
	Lucid	Original
	(1)	(2)
Rumor Only	-0.008	-0.031
	(0.044)	(0.063)
Rumor + Nonpartisan Correction	-0.018	-0.180^{***}
	(0.044)	(0.061)
Rumor + Republican Correction	-0.077^{*}	-0.175^{***}
	(0.043)	(0.062)
Rumor + Democratic Correction	-0.031	-0.186^{***}
	(0.043)	(0.061)
Constant	-0.172	-0.190
	(0.030)	(0.043)
N	3,503	1,593
R^2	0.001	0.011

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$