Research Article

Research and Politics
January-March 2019: 1–14

DOI: 10.1177/2053168018822174
journals.sagepub.com/home/rap
SAGE

Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents

Alexander Coppock[1] and Oliver A. McClellan[2]

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 in general 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 the 2012 study by Berinsky, Huber, and Lenz that compares Amazon's Mechanical Turk to US 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 US national benchmarks, with the exception of experimental treatments that aim to dispel the "death panel" rumor regarding the Affordable Care Act. We conclude that subjects recruited from the Lucid platform constitute a sample that is suitable for evaluating many social scientific theories, and can serve as a drop-in replacement for many scholars currently conducting research on Mechanical Turk or other similar platforms.

Keywords

Survey experiments, convenience samples, generalizability

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 causal effects, now they can test, refine, and retest in short order.

The main purpose of this paper is to introduce a new source of subjects – Lucid – that satisfies many desiderata, including a large subject pool, demographic diversity, and low cost. Lucid is an aggregator of survey respondents from many sources. It collects basic demographic information from all subjects who flow through their doors, facilitating quota sampling to match the US Census demographic margins. Berinsky et al. (2012) demonstrated the validity of MTurk by replicating classic experiments originally conducted on probability samples; we follow their lead and do the same on Lucid. As an empirical matter, our Lucid replications recover very similar treatment effect estimates to the original studies. That said, whether or not our particular set of replications on Lucid match previous estimates is only tangentially related to whether researchers should adopt the platform. Past success is no guarantee of future success and what worked for one experiment may not work for the next.

For this reason, a second purpose of the paper is to consider the question of when survey experimenters should opt for convenience samples in general and Lucid in particular. Our answer to this question follows the "fit-for-purpose"

[1]Yale University, New Haven, USA
[2]Columbia University, New York, USA

Corresponding author:
Alexander Coppock, 77 Prospect Street, New Haven, CT 06511, USA.
Email: alexander.coppock@yale.edu

compromise proposed by public opinion scholars in an attempt to resolve debates over non-probability samples. The basic distinction drawn in Baker et al. (2013) is between descriptive work, which requires probability samples, and work that "models relationships between variables," which can make fruitful use of non-probability samples. Similarly, we think that if the purpose of a study is to estimate sample average treatment effects (SATEs), convenience samples are usually fit for purpose. A key distinction will be whether the causal effects under study should, according to the social scientific theories guiding the design of the experiment, obtain among the convenience sample as well as a broader population. In our experience, it is the rare theory whose scope conditions specifically exclude the sort of people who take online surveys, though one could come up with counterexamples, for example theories whose predictions depend on the level of digital literacy (Koltay, 2011).

If it is determined that a convenience sample is fit for purpose for survey experimentation, the question remains which source to use. MTurk is a widely used platform and scholars know a tremendous amount about it, both positive and negative. On the positive side, recent meta-analyses of experimental studies conducted on both MTurk and US national probability samples (Coppock, 2017; Coppock et al., 2018; Mullinix et al., 2015) have found high replication rates. On the negative side, Behrend et al. (2011) show that MTurk responses are slightly more susceptible to social desirability bias than other samples. Others are concerned that MTurk 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 (De Quidt et al., 2018; Mummolo and Peterson, 2018; White et al., 2018), even when indicating the investigators' preferred responses with heavy-handed messages. Some scholars are concerned that MTurk is "overfished" and that many respondents have become professional survey takers (Chandler et al., 2015; Rand et al., 2014). Stewart et al. (2015) estimate the pool of active MTurk respondents for a given lab to be approximately 7300 subjects at any one time. Lastly, 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 late 2018, much of the academic community was shaken by the revelation that many MTurk responses were fraudulent or even "bots" (Dennis et al., 2018). While tools to circumvent the problem were very quickly produced (Ahler et al., 2018; Kennedy et al., 2018), the episode underlined the dangers inherent in overreliance on any one source of subjects.

When are convenience samples fit for purpose?

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 in general.[1] Convenience samples have met with resistance largely because they have no design-based justification for generalizing from the sample to the population and typically have to rely on some combination of statistical adjustment and argument instead. Debates over the scientific status of non-probability samples have raged for decades. Warren Mitofsky's 1989 presidential address to the American Association for Public Opinion Research (AAPOR) describes an acrimonious dispute from a half-century prior over quota versus probability sampling in which "[t]here was no meeting of the minds among the participants." In that same address, Mitofsky describes his own journey from probability-sample purist to convenience-sample convert, at least in some settings and for some scientific purposes.

In 2013, AAPOR issued a report on non-probability samples (Baker et al., 2013) that formalizes a "fit-for-purpose" framework for assessing whether a given sampling design is fit for the scientific purpose to which it is put. The fit-for-purpose framework represents a compromise: for descriptive work, we need probability samples, but for research that models the relationships between variables, convenience samples may be acceptable.[2] Levay et al. (2016) provides some empirical support for the compromise's underlying reasoning: MTurk and probability samples are descriptively quite different, but the correlations among survey responses are similar after a modicum of statistical adjustment. And 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). We would note, however, that even extremely idiosyncratic convenience samples (e.g. Xbox users; Gelman et al., 2016) can sometimes produce estimates that turn out to have been accurate. Nevertheless, in line with the fit-for-purpose framework, we would not generally recommend using Lucid (or any convenience sample) when the goal is descriptive inferences that are representative of a particular population.

In contrast to descriptive studies which seek to estimate a population quantity on the basis of a sample, the goal of much experimental work is to estimate a particular *sample*

quantity, the 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.[3]

But the question of whether a particular convenience sample should be used depends not on whether we can estimate the SATE well, but on whether the SATE is worth estimating at all. In our view, the choice to use a convenience sample should depend on whether the SATE is *relevant for theory*. A similar distinction is drawn in Druckman and Kam (2011), who were responding to the critique of student samples given in Henrich et al. (2010). Druckman and Kam (2011) point out that a convenience sample might pose a problem if it lacks variation on an important moderating variable. Indeed, variation in the moderator is required to demonstrate that effects are different for different subgroups, but we would submit that even in the absence of such variation, the SATE in a convenience sample could be 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 an 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 US 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.

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.[4] 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.

Why would SATEs and PATEs ever differ? We need to distinguish between three kinds of heterogeneity: idiosyncratic, treatment-by-covariate, and treatment-by-treatment (Gerber and Green, 2012: ch. 9). Idiosyncratic heterogeneity occurs when subjects' responses to treatment are different, but this heterogeneity is not caused by systematic factors. Treatment-by-covariate heterogeneity occurs when groups of subjects defined by pre-treatment covariates have different average responses to treatment. This kind of heterogeneity can cause SATEs and PATEs to differ if the covariates that are correlated with treatment effects are also correlated with the characteristics that influence selection into the convenience sample (Hartman et al., 2015; Kern et al., 2016). If these important moderators are measured in the sample and are known in the population, then SATEs can be reweighted to estimate PATEs (Franco et al., 2017, Miratrix et al., 2018). Lastly, treatment-by-treatment heterogeneity occurs when the response to one treatment depends on the level of another treatment, as in a two-by-two factorial design. In our empirical section, we investigate both treatment-by-covariate and treatment-by-treatment interactions.

As it happens, survey experimental SATE and PATE estimates are frequently quite similar (Coppock, 2017; Coppock et al., 2018; Mullinix et al., 2015), 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 MTurk. Whether or not future experiments will also exhibit low treatment effect heterogeneity is, of course, only a matter of speculation.

A second kind of external validity is about whether the treatments and outcomes in the experiment map on to the "real-world" treatments and outcomes that the study is meant to illuminate. This sort of external validity has less to do with who the experimental subjects happen to be and more to do with the strength of the analogy from the experimental design to the social or political phenomenon of interest. Our ability to aggregate experimental findings into a broader understanding of politics and society is arguably much more important than the relative magnitudes of particular SATEs and PATEs. Assessing this kind of external validity is outside the scope of the current paper, but our guess is that the choice of one convenience sample over another does not alter it for better or worse.

In our empirical section, we replicate five survey experiments that were originally conducted on other samples. As an exercise, we read each paper with an eye towards understanding whether the theory under study should, in principle, apply to the sorts of people who participate in online surveys. We also noted whether treatment effects were predicted to be moderated by particular variables in the original paper. This is relevant because, as noted in Druckman and Kam (2011), a sample needs sufficient variation on a moderating variable in order to demonstrate the presence of treatment effect heterogeneity. Table 1 pulls together the results of this exercise. In three cases, the group to whom the theory appears to apply is all adult English-speaking Americans and, in two cases, the groups is simply all adult humans. Lucid subjects are strict subsets of both groups. The theoretical moderators

Table 1. Theoretical applicability of five experimental studies.

Experiment	Theory	Relevant sample
GSS welfare	Subjects will be more willing to support "assistance to the poor" than they will to support "welfare." This effect may be moderated by education, ideology, and/or gender.	The adult English-speaking population of the USA.
Tversky and Kahneman (1981)	Framing trade-offs in terms of losses leads to risk aversion. Framing trade-offs in terms of gains leads to risk tolerance. This framing effect is assumed to operate similarly for all individuals.	All adult humans.
Kam and Simas (2010)	Subjects' baseline risk acceptance may condition the magnitude of Tversky and Kahneman's (1981) framing effects.	All adult humans.
Hiscox (2006)	Framing effects can influence how subjects evaluate free-trade policies. These framing effects may be moderated by education.	The adult English-speaking population of the USA.
Berinsky (2017)	Correcting misinformation can lead to opinion reversal, depending on the source the correction comes from. Treatment effect may vary by subject attentiveness as well as partisanship.	The adult English-speaking population of the USA.

were education, ideology, gender, risk acceptance, education, subject attentiveness, and partisanship.

Our sample

In this section, we describe the Lucid platform, how subjects are recruited, and the distributions of their demographic, psychological, and political attributes.

Subject recruitment

Lucid is the largest marketplace for online "sample" in the USA. 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 three months), their demographic characteristics (age, gender, ethnicity, race, education, income, and ZIP code) are measured using the US 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.[5] Lucid can construct demographically targeted sets of respondents using a combination of quota sampling and screening questions. For example, Flores and Coppock (2018) obtained 2866 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 US$1 per completed response as of 2018. See Graham (2018) for a Lucid sample constructed in this manner.

Just like MTurk (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. This raises concerns about "production transparency" (Journal Editors' Transparency Statement, 2014). As we only have a single sample of 3504 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. 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 divergent results in one of our original–replication pairs.

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. Respondents report taking an average of 4.28 surveys per month. However, 98% 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 US$5.01, but if we trim off responses that are implausible (greater than US$20.00), we obtain the more reasonable figure of US$1.16.

Baseline characteristics

Before comparing the SATEs obtained on Lucid to those obtained on MTurk and on probability samples, we assess

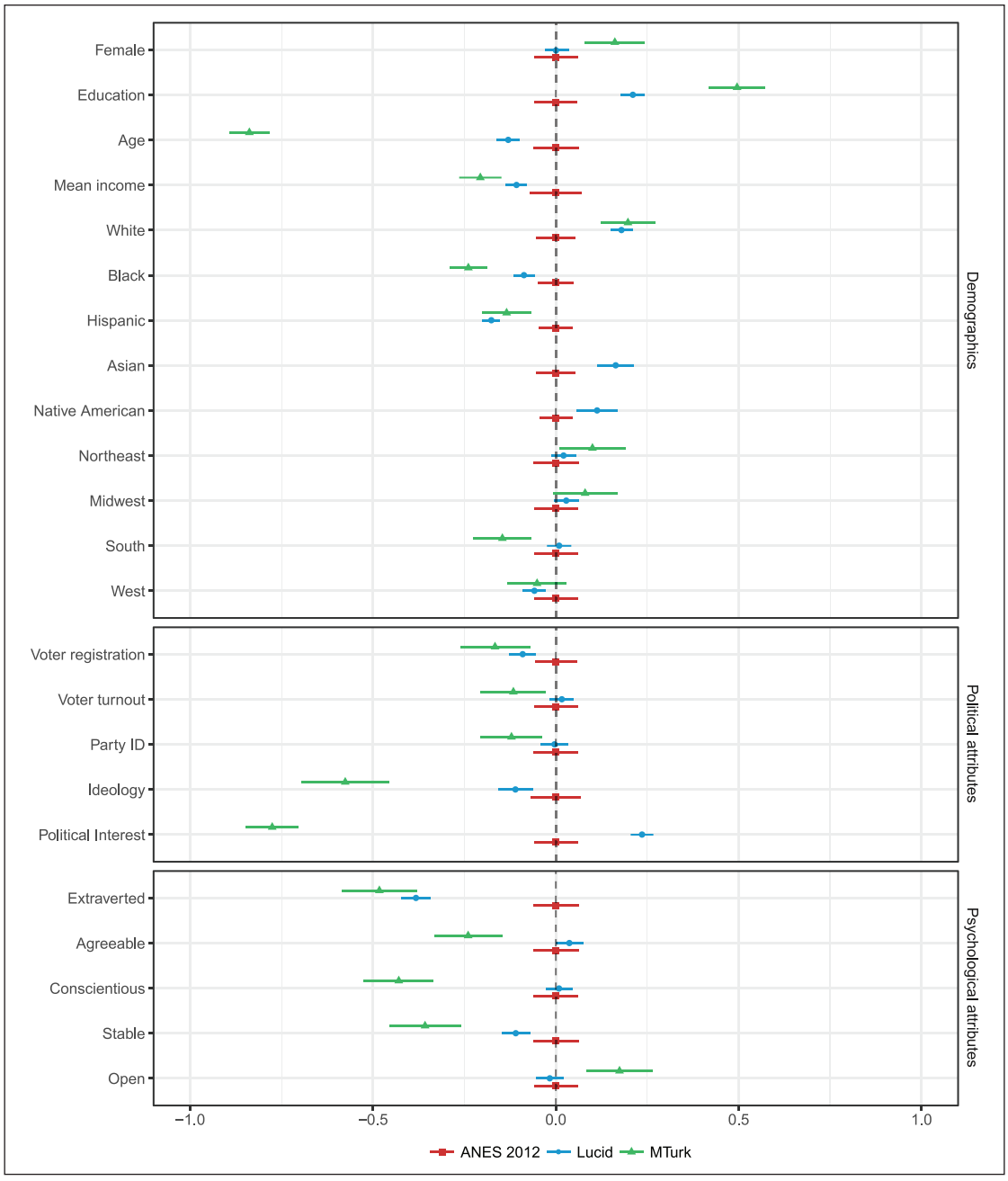

Figure 1. Standardized means for demographic variables.

the distribution of baseline characteristics like demographics, political attitudes, and psychological traits. Figure 1 presents standardized demographic means on Lucid, MTurk, and the 2012 American National Election Study (ANES), where we standardize by the ANES 2012 mean and standard deviation.[6]

In terms of gender, education, age, and income, the Lucid sample comes closer to the ANES 2012 benchmarks than the MTurk sample does. 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 non-whites, 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 nine instances, five of which are statistically significant.[7]

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.[8] 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 five-point political interest scale. This trend is reversed for political knowledge, included in the online appendix, Table 2. MTurk respondents scored higher on political knowledge 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 the online appendix, 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.

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 Cooperative Congressional Election Study (CCES), Cooperative Campaign Analysis Project (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 US national samples on psychological traits. Formal hypothesis tests demonstrate that Lucid is significantly closer to the ANES 2012 than MTurk on all five traits.

Experiments

We now turn to our five replication experiments. For space reasons, we provide brief descriptions of each experiment in the main text along with summary figures 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, in the main, we follow the analysis strategies of the original authors. Again, following the original authors, we drop subjects with missing or don't know outcomes.[9] 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 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

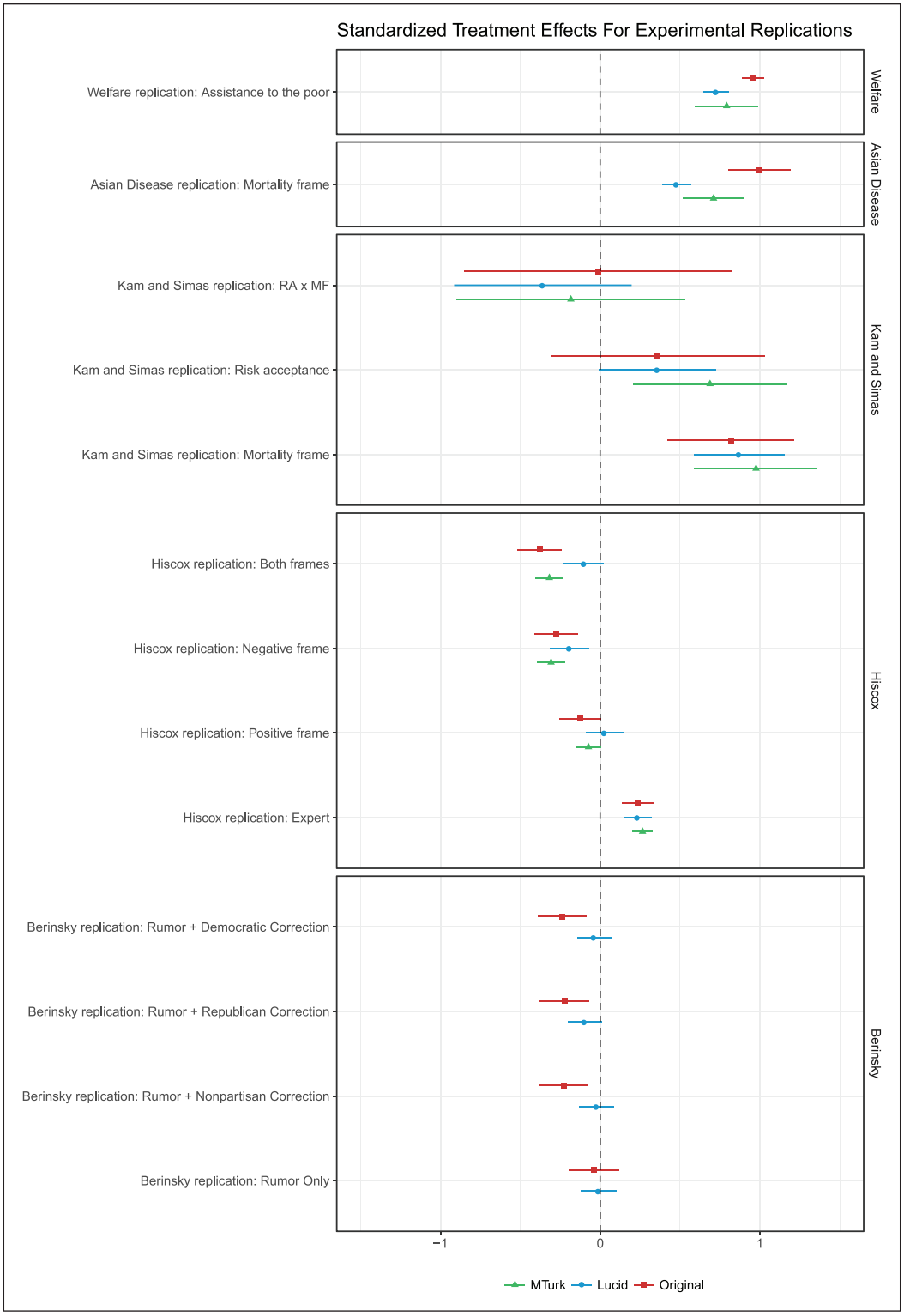

Figure 2. Summary of experimental comparisons across samples.

sample. Lacking a US 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.

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

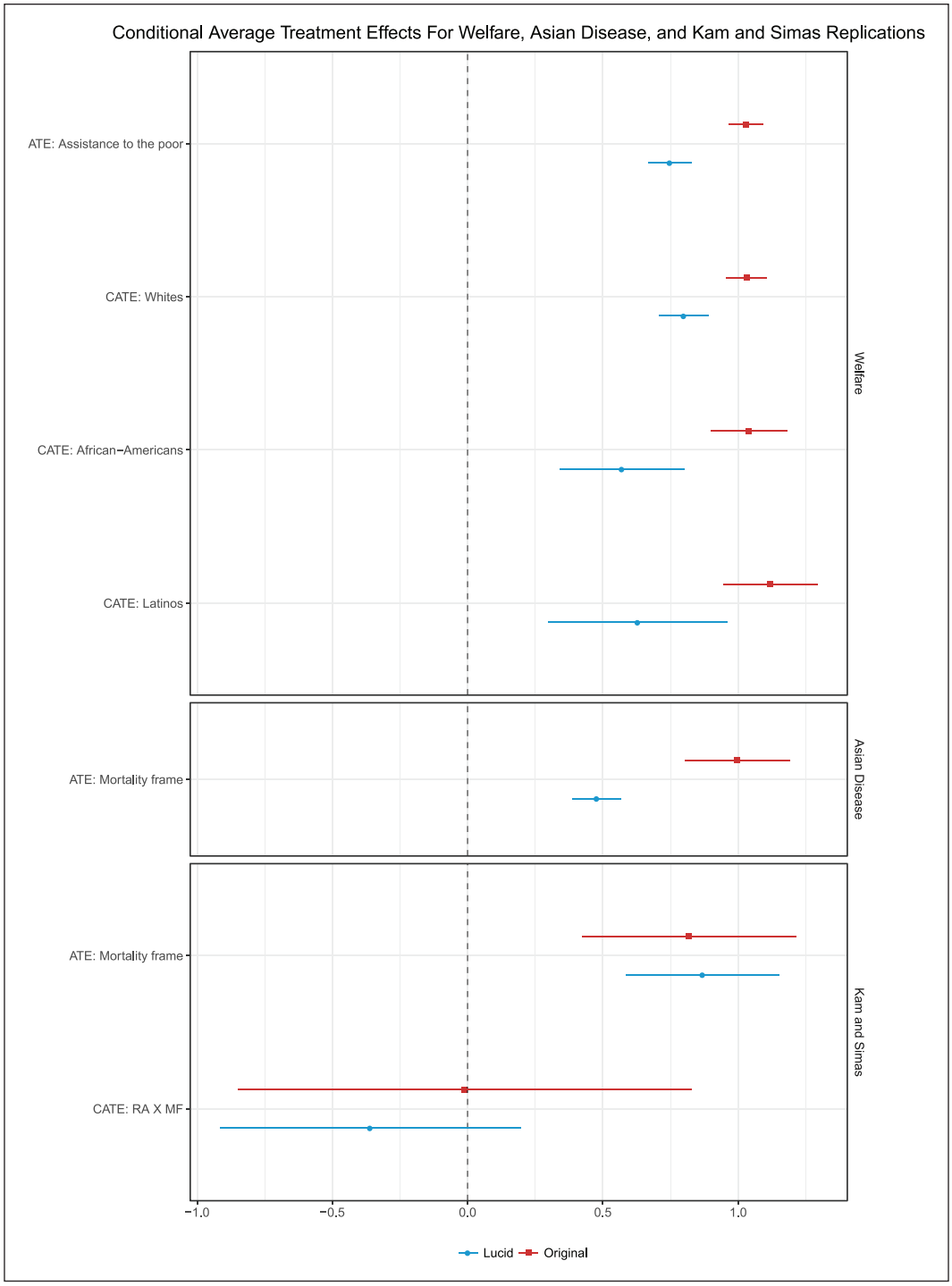

Figure 3. Treatment effect heterogeneity for welfare spending, Asian Disease, and Kam and Simas (2010).
RA: risk acceptance covariate, MF: mortality frame treatment.

treatment effect of the mortality frame does not vary appreciably with risk acceptance.[10] 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, as we will discuss in greater depth.

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 × 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: Healthcare 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 (ACA), 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 (SSI)), a large portion of the sample believed the rumor, and corrections delivered by Republicans, Democrats, and Nonpartisan groups were all 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 belief were −0.17 on Lucid, compared with −0.19 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 the 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.

Treatment effect heterogeneity

As previously discussed, in addition to estimating average treatment effects (ATEs) for overall sample populations, one of the important determinants of whether an experimental sample is fit for purpose is whether the sample can be used to estimate conditional average treatment effects (CATEs). In this section, we assess treatment effect heterogeneity in four of the five experiments replicated above.[11]

For the welfare spending experiment, we test whether subjects' race or ethnicity conditions the effect of the "assistance to the poor" phrasing. Though this is not one of the factors theorized to condition the phrasing treatment effect in the original iteration of this experiment, race has since been identified as perhaps the single most important influencing factor in position on welfare spending (Gilens, 1996). We assess whether the treatment effect of receiving the "assistance to the poor" versus "welfare" phrasing varies among white, black, and Latino respondents, among both the Lucid sample and respondents in the 2016 GSS. Both samples generally exhibit low treatment effect heterogeneity, with CATEs among white, black, and Latino respondents being statistically indistinguishable from one another. While the point estimates for the 2016 GSS CATEs are nearly identical, in the Lucid sample white respondents exhibited larger treatment effect magnitude than black or Latino respondents, though again these differences are not statistically significant. These estimates are shown in the first facet of Figure 3.

In the third facet of Figure 3, we assess whether subjects' prior risk acceptance conditions the effect of receiving the mortality frame. In neither sample do we see a significant conditioning effect for risk assessment – both the original sample and Lucid sample are able to replicate estimates of (the lack of) heterogeneous treatment effects.

For the Hiscox free-trade framing experiment, we test for two different types of treatment effect heterogeneity, as can be seen in Figure 4. We assess both heterogeneity based on respondents' prior characteristics, in the form of education levels, as well as heterogeneity that is randomly assigned as part of the experimental design, in whether or not subjects receive the summary of expert opinions. Across both the original sample and Lucid sample, we see no evidence of heterogeneous treatment effects for any of the possible treatment conditions. While the results seem to suggest that subjects with "low" education levels, as defined in Hiscox (2006) as subjects who have not attended any college, are slightly more influenced by both framing and expert opinions, these differences are far from statistically significant. Receiving the expert opinions does not appear to moderate the effect of receiving any of the possible treatment frames.

Treatment effect heterogeneity for the healthcare rumors is also low for both the original Berinsky (2017) sample and for the Lucid sample (Figure 5). Treatment effects for all possible treatment possibilities are similar for both Democrats and Republicans for both samples. It is important to note that while we do not replicate the original ATEs for this study in the previous section, here we see that the CATEs are statistically differentiable for only Democrats receiving the Democratic correction to the healthcare rumor. We can, therefore, clarify our findings for this replication. Subjects identifying as Democrats sampled from SSI in 2010 reacted differently to the Democratic correction than did Democratic subjects sampled from Lucid in 2016. Whether this difference is due to altered political context over time, solidification of beliefs towards the ACA, or differences in the sample pools between SSI and Lucid cannot be stated for certain.

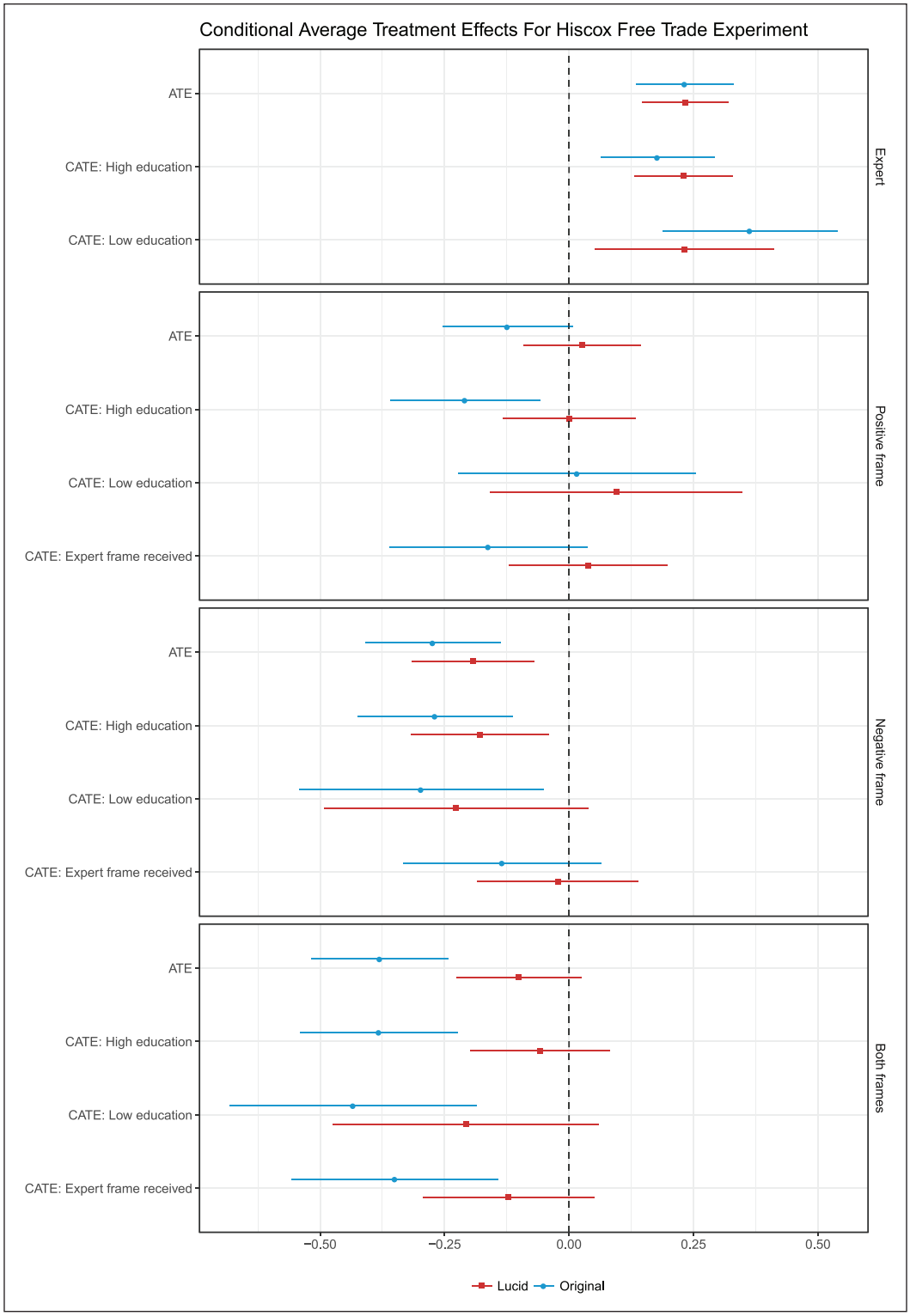

Figure 4. Treatment effect heterogeneity for Hiscox free trade.

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

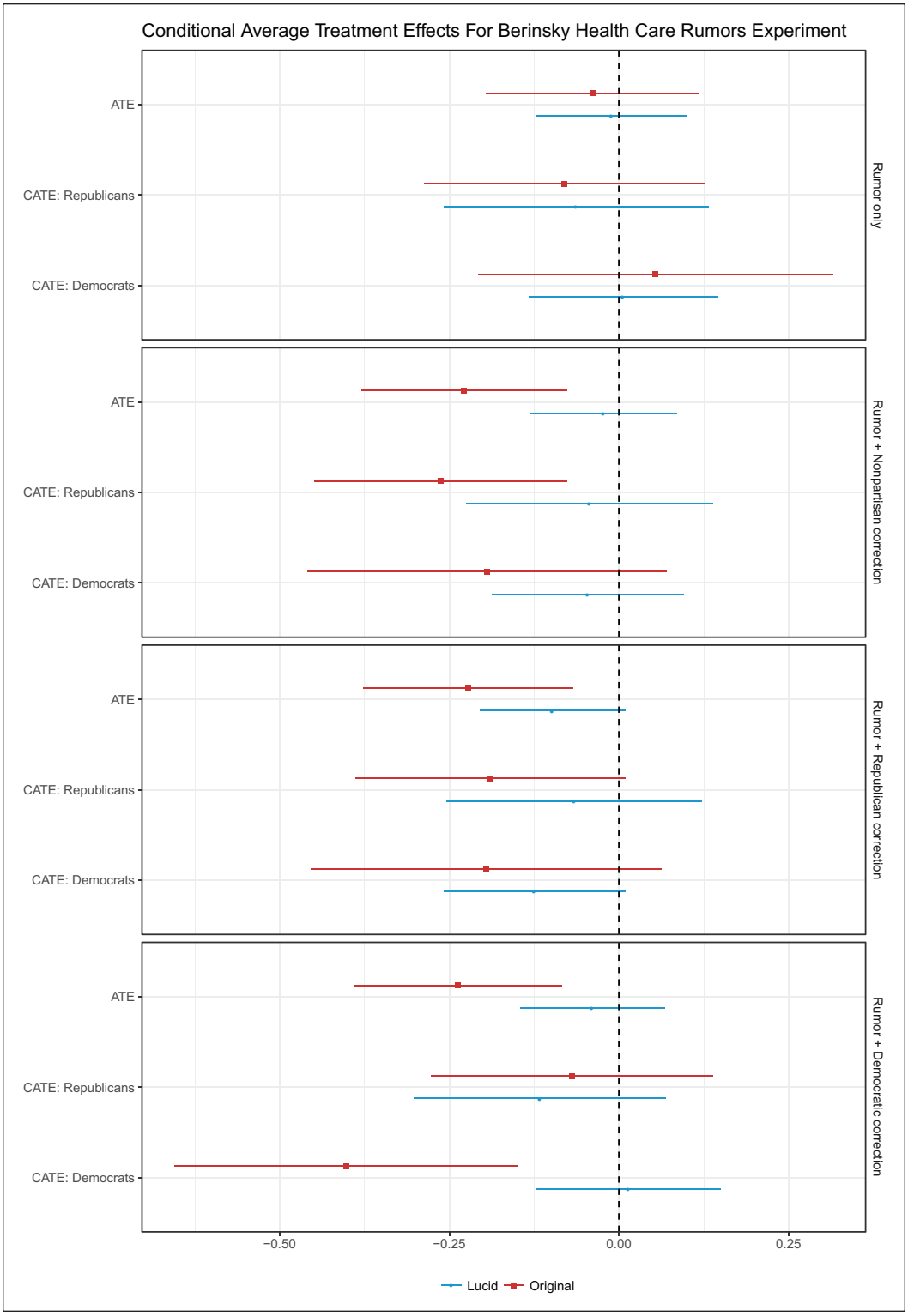

Figure 5. Treatment effect heterogeneity for Berinsky healthcare rumors.

samples appropriate? Second, when convenience samples are appropriate, is MTurk the best option, or are there alternatives?

We have relied on the fit-for-purpose framework to answer the first question. The purpose of most survey experiments is to estimate a SATE; whether a given SATE is interesting depends on whether the sample is relevant for theory. Theoretical relevance concerns whether the theory's predictions extend to a particular sample, not whether the

sample is drawn at random from some population. While we think the theories that underlie most survey experiments conducted in the USA would extend to Lucid, we emphatically do not mean to that any experiment conducted on a convenience sample is relevant for theory.

In our five experiments, Lucid performed remarkably well in recovering estimates that come 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. We test this heterogeneity directly and conclude that in nearly every case, low treatment effect heterogeneity is indeed the reality, at least along the dimensions we assess.

Among our five experiments, we have one instance of the Lucid sample producing substantively different results compared to the original study. 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 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. Alternatively, we might say that, ex-ante, we considered the Lucid sample relevant for theory and these new results require us to update the theory forwarded in that paper.

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 MTurk. Lucid boasts a much larger pool of subjects than MTurk; the risk of cooperation among subjects is minimal given their diverse sources; subjects are less professionalized; subjects are more similar to US 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. That said, researchers have developed tools to implement a wide variety of studies on MTurk. For example, the MTurk 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 MTurk 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 MTurk data, we also recognize that changing data sources does not come without costs.

Authors' note

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). Replication code and data are available here: https://doi.org/10.7910/DVN/DDWWJW.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental materials

The supplemental files are available at http://journals.sagepub.com/doi/suppl/10.1177/2053168018822174.
The replication files are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DDWWJW

Notes

1. Sometimes researchers employ people encountered online (through MTurk 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. We note in passing that the distinction drawn in Baker et al. (2013) is (possibly purposely) vague. "Modeling relationships between variables" might also reasonably be considered descriptive research requiring probability samples if the quantity to be described is, for example, the population covariance between two variables. As the focus of this paper is exclusively on causal estimands, or the relationship between treatments and outcomes, this ambiguity is not of great concern here.
3. Gerber and Green (2012: ch. 2) for a discussion of the three core assumptions required for internally valid inference in an experiment.
4. 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.

5. These figures were 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.
6. Corresponding tables that include comparisons to additional samples are shown in the online appendix. 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 poststratification weights from Wave 11, the latest wave from which we analyze data.
7. See the online appendix for the bootstrapping procedure we used to conduct these tests.
8. All surveys examined here significantly overstate both turnout and registration compared to official government statistics. For a detailed discussion of this phenomenon, see DeBell et al. (2018).
9. 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: ch. 7).
10. 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 ATE (or, as in this case, the CATEs). See Gerber and Green (2012: ch. 2) for a textbook proof. The substantive conclusions do not change in any way if we use probit regression.
11. We omit Tversky and Kahneman's (1981) Asian Disease experiment from this heterogeneity test, as the Kam and Simas (2010) experiment is itself a test of treatment effect heterogeneity in the original Asian Disease experiment.

ORCID iD

Alexander Coppock https://orcid.org/0000-0002-5733-2386

References

Ahler DJ, Roush CE and Sood G (2018) The micro-task market for "Lemons": Collecting data on Amazon's Mechanical Turk. Working Paper. Epub ahead of print.

Baker RJ, Brick M, Bates NA, et al. (2013) Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* 1(2): 90–143.

Behrend TS, Sharek DJ, Meade AW, et al. (2011) The viability of crowdsourcing for survey research. *Behavior Research Methods* 43(3): 800–813.

Berinsky AJ (2017) Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science* 47(2): 241–262.

Berinsky AJ, Huber GA and Lenz GS (2012) Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* 20(3): 351–368.

Boas TC, Christenson DP and Glick DM (n.d.) Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. *Political Science Research and Methods*. Epub ahead of print 08 August 2018. Available at: https://doi.org/10.1017/psrm.2018.28

Bullock JG, Gerber AS, Hill SJ, et al. (2015) Partisan bias in factual beliefs about politics. *Quarterly Journal of Political Science* 10(4): 519–578.

Chandler J, Paolacci G, Peer E, et al. (2015) Using nonnaive participants can reduce effect sizes. *Psychological Science* 26(7): 1131–1139.

Coppock A (2017) Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*. Epub ahead of print 27 March 2018. Available at: https://doi.org/10.1017/psrm.2018.10.

Coppock A, Leeper TJ and Mullinix KJ (December, 2018) Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences* 115(49): 12441–12446. DOI:10.1073/pnas.1808083115

DeBell M, Krosnick JA, Gera K, et al. (2018) The turnout gap in surveys: Explanations and solutions. *Sociological Methods & Research*. Epub ahead of print 6 May 2018. Available at: https://doi.org/10.1177/0049124118769085

De Quidt J, Haushofer J and Roth C (2018) Measuring and bounding experimenter demand. *American Economic Review* 108(11): 3266–3302. DOI: 10.1257/aer.20171330

Dennis SA, Goodson BM and Pearson C (2018) MTurk workers' use of low-cost "virtual private servers" to circumvent screening methods: A research note. Working Paper. Available at SSRN: https://ssrn.com/abstract=3233954 or http://dx.doi.org/10.2139/ssrn.3233954

Druckman JN and Kam CD (2011) Students as experimental participants: A defense of the "Narrow Data Base". In: Druckman JN, Green DP, Kuklinski JH, et al. (eds) *Cambridge Handbook of Experimental Political Science*. Cambridge: Cambridge University Press, pp.41–57.

Flores A and Coppock A (2018) Do bilinguals respond more favorably to candidate advertisements in English or in Spanish? *Political Communication* 35(4): 612–633. DOI: 10.1080/10584609.2018.1426663

Franco A, Malhotra N, Simonovits G, et al. (2017) Developing standards for post-hoc weighting in population-based survey experiments. *Journal of Experimental Political Science* 4(2): 161–172.

Gelman A, Goel S, Rivers D, et al. (2016) The mythical swing voter. *Quarterly Journal of Political Science* 11(1): 103–130.

Gerber AS and Green DP (2012) *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

Gerber AS, Huber GA, Doherty D, et al. (2010) Personality and political attitudes: Relationships across issue domains and political contexts. *American Political Science Review* 104(01): 111–133.

Gilens M (1996) "Race coding" and white opposition to welfare. *American Political Science Review* 90(3): 593–604.

Gosling SD, Rentfrow PJ and Swann WB Jr (2003) A very brief measure of the big-five personality domains. *Journal of Research in Personality* 37(6): 504–528.

Graham MH (2018) Self-awareness of political knowledge. *Political Behavior*. Epub ahead of print 08 September 2018. Available at: https://doi.org/10.1007/s11109-018-9499-8

Hartman E, Grieve R, Ramsahai R, et al. (2015) From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(3): 757–778.

Henrich J, Heine SJ and Norenzayan A (2010) The weirdest people in the world? *Behavioral and Brain Sciences* 33(2–3): 61–83. DOI: 10.1017/S0140525X0999152X

Hiscox MJ (2006) Through a glass and darkly: Attitudes toward international trade and the curious effects of issue framing. *International Organization* 60(03): 755–780.

Journal Editors' Transparency Statement (2014) Available at: https://www.dartstatement.org/2014-journal-editors-statement-jets (accessed 05 January 2019).

Kam CD and Simas EN (2010) Risk orientations and policy frames. *The Journal of Politics* 72(2): 381–396.

Kennedy R, Clifford S, Burleigh T, et al. (October, 2018) The shape of and solutions to the MTurk quality crisis. Available at SSRN: https://ssrn.com/abstract=3272468 or http://dx.doi.org/10.2139/ssrn.3272468

Kent ML, Harrison TR and Taylor M (2006) A critique of internet polls as symbolic representation and pseudo-events. *Communication Studies* 57(3): 299–315.

Kern HL, Stuart EA, Hill J, et al. (2016) Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness* 9(1): 103–127.

Koltay T (2011) The media and the literacies: Media literacy, information literacy, digital literacy. *Media, Culture & Society* 33(2): 211–221.

Leeper TJ (2015) *MTurkR: Access to Amazon Mechanical Turk Requester API via R*. R package, version 0.6.5.1.

Levay KE, Freese J and Druckman JN (2016) The demographic and political composition of Mechanical Turk samples. *SAGE Open* 6(1): 1–17.

Mason W and Suri S (2012) Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44(1): 1–23.

Miratrix LW, Sekhon JS, Theodoridis AG, et al. (2018) Worth weighting? How to think about and use weights in survey experiments. *Political Analysis* 26(3): 1–17.

Mitofsky WJ (1989) Presidential address: Methods and standards: A challenge for change. *Public Opinion Quarterly* 53(3): 446–453.

Mullinix KJ, Leeper TJ, Druckman JN, et al. (2015) The generalizability of survey experiments. *Journal of Experimental Political Science* 2(2): 109–138.

Mummolo J and Peterson E (2018) Demand effects in survey experiments: An empirical assessment. *American Political Science Review*. Epub ahead of print 11 December 2018. Available at: https://doi.org/10.1017/S0003055418000837

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

Park DK, Gelman A and Bafumi J (2004) Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis* 12(4): 375–385.

Rand DG, Peysakhovich A, Kraft-Todd GT, et al. (2014) Social heuristics shape intuitive cooperation. *Nature Communications* 5. Available at: https://www.nature.com/articles/ncomms4677

Stewart N, Ungemach C, Harris AJL, et al. (2015) The average laboratory samples a population of 7,300 Amazon Mechanical Turk Workers. *Judgment and Decision Making* 10(5): 479.

Tversky A and Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481): 453–458.

White A, Strezhnev A, Lucas C, et al. (2018) Investigator characteristics and respondent behavior in online surveys. *Journal of Experimental Political Science* 5(1): 56–67.