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

Alexander Coppock and Oliver A. McClellan*

August 30, 2017
Draft prepared for presentation at the 113th Annual Meeting of the American Political Science Association in San Francisco, CA, August 31 - September 3, 2017.

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 that they have become "overfished" in the sense that many participants on these platforms have become professional survey takers. 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.

Concerns over the use of MTurk and similar online 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 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 (e.g., Mullinix et al., 2015) have found generally high degrees of correspondence across platforms, successes in certain domains do not fully alleviate these concerns.

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.

Finally, and most importantly for this article, some scholars are concerned that MTurk is "overfished" and that many respondents have become professional survey takers (Rand et al., 2014; Chandler et al., in press). 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. This behavior is particularly troubling for surveys where subjects' payoffs can depend 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).

One response to these concerns has been to downweight evidence that relies on MTurk data (Kahan, 2013). Another response has been to replicate MTurk studies on alternative platforms to assess empirically the extent to which conclusions depend on sample selection (Mullinix et al., 2015). Each of these responses require new sources of survey subjects. Well-known alternatives such as YouGov, GfK, and SSI are often much more expensive per response than Mechanical Turk.

An alternative to MTurk should satisfy four major desiderata. It should feature a larger pool of respondents, respondents should not to be able to coordinate, samples should better approximate target populations, and, it should be competitive in terms of cost.

In this paper, we demonstrate the viability of Lucid's Fulcrum Exchange as a source of survey subjects for the social sciences. Lucid is the largest marketplace for online "sample" nationwide. Providers direct respondents to Lucid, which then redirects subjects to purchasers, typically market research firms. Approximately 350,000 unique respondents (estimated from unique IP addresses) pass through the exchange each day; in 2015, Lucid managed 30 million unique respondents (figures

obtained via private correspondence). Unlike MTurk, but similar to YouGov, GfK, and SSI, Lucid samples can be demographically targeted.[1] For a 10-minute survey delivered to a group of subjects quota sampled to match census demographics, researchers can expect to pay approximately 1 dollar per completed response as of 2017.

We replicate the suite of studies presented in Berinsky et al. (2012) on Lucid. Berinsky et al. (2012) compare Mechanical Turk samples to other convenience and probability samples in terms of demographics, political attitudes, and three brief survey experiments. To these analyses, we add cross-sample comparisons of the Big Five personality traits and two additional survey experiments first reported in Berinsky (forthcoming) and Hiscox (2006). To preview our results, we find that Lucid generally outperforms MTurk in terms of recovering estimates obtained on probability samples. With some exceptions (detailed below), we find a strong correspondence between Lucid estimates and national benchmark targets.

1 Baseline Characteristics

We first compare our Lucid sample to other samples in terms of baseline characteristics like demographics, political attitudes, and psychological traits.

1.1 Demographics

Table 1 compares the demographic profiles of six surveys; three internet samples (Lucid, an MTurk sample drawn in 2010 and reported Berinsky et al. (2012), and the American National Election Survey (ANES) Panel study) and three face-to-face samples (the 2008 CPS and the 2008 and 2012 ANES). Where survey weights were provided, all entries in the table represent weighted means. The Lucid sample is unweighted.

In terms of gender, the Lucid sample comes closest of any of the six surveys to the 2010 US census value of 50.8% female. The mean number of years of education on Lucid (14.2) is higher than the approximately 13.5 years recorded by the face-to-face surveys, but is closer than the other two internet sample estimates. Both mean and median incomes are lower on Lucid than among the face-to-face samples, but are higher than in the MTurk sample. All of the internet samples overrepresent whites relative to nonwhites, but this distortion is smallest on Lucid. The regional balance on Lucid comes very close to the 2008 CPS, whereas the MTurk sample appears to overrepresent southerners.

1.2 Politics

Table 2 contains a summary of the political and psychological makeup of the examined samples. Voter registration and turnout seem to vary somewhat across samples, with the ANES Panel

[1] For example, Flores and Coppock (2016) targeted 2,866 bilingual subjects using a custom screening question.

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)	57.58 (0.90)	51.67 (0.18)	54.98 (1.30)	52.01 (1.53)
Education (mean years)	14.16 (0.04)	14.88 (0.10)	16.22 (0.06)	13.21 (0.01)	13.49 (0.06)	13.63 (0.07)
Age (mean years)	44.92 (0.28)	32.26 (0.49)	49.71 (0.27)	46.02 (0.06)	46.55 (0.47)	47.25 (0.57)
Mean income	$60,896.33 ($833.41)	$55,331.82 ($1,659.29)	$71,761.60 ($897.47)	$62,255.92 ($149.25)	$66,424.82 ($1,437.13)	$66,948.92 ($2,039.72)
Median income	$47,500	$45,000	$67,500	$55,000	$55,000	$52,500
Race						
White (%)	78.87 (0.69)	79.67 (1.72)	81.70 (0.71)	68.50 (0.17)	74.85 (0.93)	70.70 (1.24)
Black (%)	9.11 (0.49)	4.17 (0.85)	8.75 (0.52)	11.67 (0.12)	12.09 (0.59)	11.93 (0.81)
Hispanic (%)	5.41 (0.38)	6.72 (1.07)	5.04 (0.40)	13.66 (0.13)	7.87 (0.43)	10.92 (0.73)
Asian (%)	3.44 (0.31)		2.64 (0.29)	5.02 (0.08)	2.27 (0.43)	1.47 (0.33)
Native American (%)	1.03 (0.17)		0.97 (0.18)	1.07 (0.04)	1.28 (0.30)	0.36 (0.14)
Region of the United States						
Northeast (%)	18.92 (0.66)	21.96 (1.77)	16.90 (0.72)	18.42 (0.14)	14.57 (1.00)	18.10 (1.23)
Midwest (%)	23.79 (0.72)	25.95 (1.87)	28.30 (0.86)	21.91 (0.14)	21.18 (1.12)	22.60 (1.27)
South (%)	37.63 (0.82)	30.13 (1.96)	31.38 (0.89)	36.54 (0.18)	42.84 (1.28)	37.20 (1.46)
West (%)	19.66 (0.67)	19.96 (1.70)	23.42 (0.81)	23.13 (0.15)	21.41 (0.99)	22.10 (1.26)
N	3,504	551	3,049	100,008[1]	2,323	2,054

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

showing the highest rates of turnout and registration and the CPS showing the lowest. All samples seem to share similar characteristics with regard to political party affiliation; the mean on the standard 7-point scale is between 3.5 and 3.9 for all datasets. We do see some important variation with regard to respondents' ideologies, however. Respondents on MTurk are markedly more liberal than respondents in the other four samples. Though Lucid respondents seem to be slightly more liberal than respondents found in the ANES Panel or face-to-face surveys, this difference is much smaller than the gap between MTurk respondents and the rest of the samples.

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. MTurk respondents scored higher on political knowledge then did respondents on the ANES Panel while Lucid respondents scored lower; gaps in both directions average about six percentage points. We speculate that this discrepancy 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.

The average policy preferences held by each of the samples in a variety of domains are shown in Table 4. 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.

1.3 Psychology

Table 3 compares samples on four commonly used psychological measures. To measure the "Big 5" personality features, we use the Ten Item Personality Inventory (Gosling et al., 2003), which

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)	92.01 (0.68)	71.00 (0.17)	78.19 (1.08)	84.75 (1.05)
Voter turnout (%)	76.33 (0.72)	70.64 (1.95)	89.83 (0.58)	63.64 (0.18)	70.43 (1.19)	70.18 (1.39)
Traits						
Party identification (mean on 7-point scale, 7= Strong Republican)	3.73 (0.04)	3.49 (0.09)	3.90 (0.05)		3.72 (0.05)	3.73 (0.06)
Ideology (mean on 7-point scale, 7= Strong conservative)	4.09 (0.04)	3.39 (0.09)	4.31 (0.05)		4.24 (0.05)	4.25 (0.05)
Political Interest[1](mean on 5-point scale, 5 = Extremely interested)	3.62 (0.02)	2.43 (0.04)	3.29 (0.02)		2.94 (0.04)	3.34 (0.04)
Political knowledge	58.39 (0.46)	70.51 (1.03)	65.59 (0.60)			
Opinions						
Prescription drug benefits for seniors (% favor)	74.19 (0.74)	63.52 (2.05)	74.78 (1.08)		80.07 (1.53)	
Universal Healthcare (% favor)	49.69 (0.84)	47.73 (2.13)	41.72 (1.23)		50.98 (1.87)	
Citizenship process for illegal immigrants (% favor)	35.99 (0.81)	38.11 (2.07)	42.67 (1.23)		49.08 (1.86)	
Ban gay marriage (% favor)	28.09 (0.76)	15.61 (1.55)	30.65 (1.15)			
Increase taxes on the rich(% favor)	63.67 (0.81)	61.16 (2.08)	55.38 (1.24)			
Increase taxes on the poor (% favor)	11.42 (0.54)	6.17 (1.03)	7.05 (0.64)			
N	3,504	551	3,049	100,008	2,323	2,054

[1] ANES 2012 uses alternate wording for political interest; see appendix for exact wording.
Entries are unweighted for Lucid, MTurk, and ANESP and are weighted for CPS 2008, ANES 2008 and ANES 2012.
Standard errors in parentheses where applicable.

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. The risk battery used by Kam and Simas (2010) evaluates how risk-acceptant subjects are. Both MTurk (0.51) and Lucid (0.49) participants appear to be mildly more risk-acceptant than Kam and Simas (2010)'s sample (0.45). The psychological measures "need to evaluate" and "need for cognition" assess the degree to which respondents are inclined towards mental stimulation and critical thinking (Cacioppo and Petty (1982), Jarvis and Petty (1996)). These measures are fairly consistent across Lucid, ANES Panel and the ANES, with MTurk coming in slightly higher. Overall, none of these samples appears to differ dramatically with respect to the average scores on commonly used psychological batteries.

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)			0.61 (0.01)	0.56 (0.01)		
Need to evaluate	0.58 (0.00)	0.63 (0.01)			0.58 (0.00)	0.56 (0.01)		
N	3,504	551[2]	1,500	20,000	3,049	2,323	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, Kam and Simas (2010) and ANESP and are weighted for CPS 2008, ANES 2008, ANES 2012, CCES and CCAP.
Standard errors in parentheses where applicable.

2 Experiments

Thus far, we have compared the performance of Lucid and MTurk with respect to baseline levels of their demographic, political, and psychological profiles. However, a large portion of academic research conducted on MTurk is experimental. The key question for survey experimenters concerned by the external validity of their results is the extent to which the sample average treatment effect (SATE) obtained on sample accords with other targets of inference, such as the SATE in another sample or the population average treatment effect (PATE) for some well-defined population.

We replicated five separate survey experiments originally conducted on other platforms on our Lucid sample. For space reasons, we provide brief descriptions of each experiment in the main text. Fuller descriptions of our procedures (including treatment and outcome question wordings) are available in the online appendix. We did not pre-register our analyses because we follow the analysis strategies of the original authors.

2.1 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 include these results for comparison. Table 4 shows that 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."

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.

2.2 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." As shown in Table 5, subjects are far more likely to choose the probabilistic (risky) outcome in when they are in the mortality (loss) frame.

Across all three samples (the original experiment was conducted in a classroom setting among undergraduates), the treatment has average effects in the same direction, 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.

2.3 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

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.

of the mortality frame does not vary appreciably with risk acceptance. Table 6 shows that those findings were replicated in both MTurk and Lucid samples. Receiving the mortality frame increases the likelihood of selecting the probabilistic choice (columns 1, 4, and 7). Risk acceptance correlates with choosing the risky option (columns 2, 5, and 8), but does not moderate the effect of treatment (columns 3, 6, and 9). This replication shows that Lucid samples are also able to replicate estimates of (the lack of) heterogeneous treatment effects.

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.

2.4 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 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?"

Table 7 shows the estimated treatment effects for four samples: Lucid, MTurk, Gfk, and the Hiscox original[2]. In all four instances of the experiment, the Expert frame increases support, the positive frame has negligible (or even negative) effects and the negative frame has unambiguously negative effects. For all three samples except Lucid, the combination of the positive and negative frames decreased support.Overall, the four studies yield similar experimental estimates.

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.

2.5 Experiment 5: Health Care Rumors

We conclude our set of five experiments with a note of caution. We attempted to replicate Berinsky's forthcoming experiment on belief in rumors surrounding the Affordable Care Act, specifically the

[2]The MTurk and GfK studies were first reported in Coppock (2016)

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), 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 belief were -0.17 on Lucid, compared with -.19 in the original. As shown in Table 8, however, none of the corrections (with the possible exception of the Republican correction) appear to have had effects anywhere near 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 constructed by SSI, a provider of online convenience samples 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.

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

3 Summary of Cross-Sample Comparisons

The foregoing results have shown that across a variety of demographic, behavioral and experimental metrics, Lucid performs at least as well as does MTurk when comparing to benchmark estimates

obtained from representative samples. Figures 1 and 2 summarize our results in graphical form. In the Demographics, Political Attributes, and Psychological Attributes facets, we present standardized means, where we standardize by the mean and standard deviation of the 2012 ANES. In the Welfare, Asian Disease, Kam and Simas, and Hiscox facets, we present standardized treatment effect estimates, where we have standardized the outcome variables by the mean and standard deviation of the original experiment.

For demographics, political behavior and psychological attributes, averages for the Lucid sample are consistently closer to those gathered in the 2012 ANES than are those gathered on MTurk. While experimental effect estimates show more variation across all sample sources, replications performed on both Lucid and MTurk produce substantively similar results compared to the original studies.

4 Survey Taking Behavior

Finally, because much of the concern over the use of MTurk has been the professionalization of subjects on the plaform, we turn to the survey-taking behavior on Lucid. As shown in Table 4, our subjects participate in a fair number of surveys. 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.

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 responses are cheap compared to those collected by alternative means, researchers can more easily conduct experiments at scale. Online surveys have also lowered the barriers to entry for early-career scholars. However, we should be cognizant of the threats to inference that may accompany an over-reliance on Mechanical Turk. Subjects on MTurk may differ from target populations in attitudes and behavior, especially as the proportion of professional survey takers increases. We think it is in our collective self-interest as scholars to diversify our database of survey responses.

In this paper, we have shown how 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

Figure 1: Summary of Baseline Characteristic Comparisons Across Samples

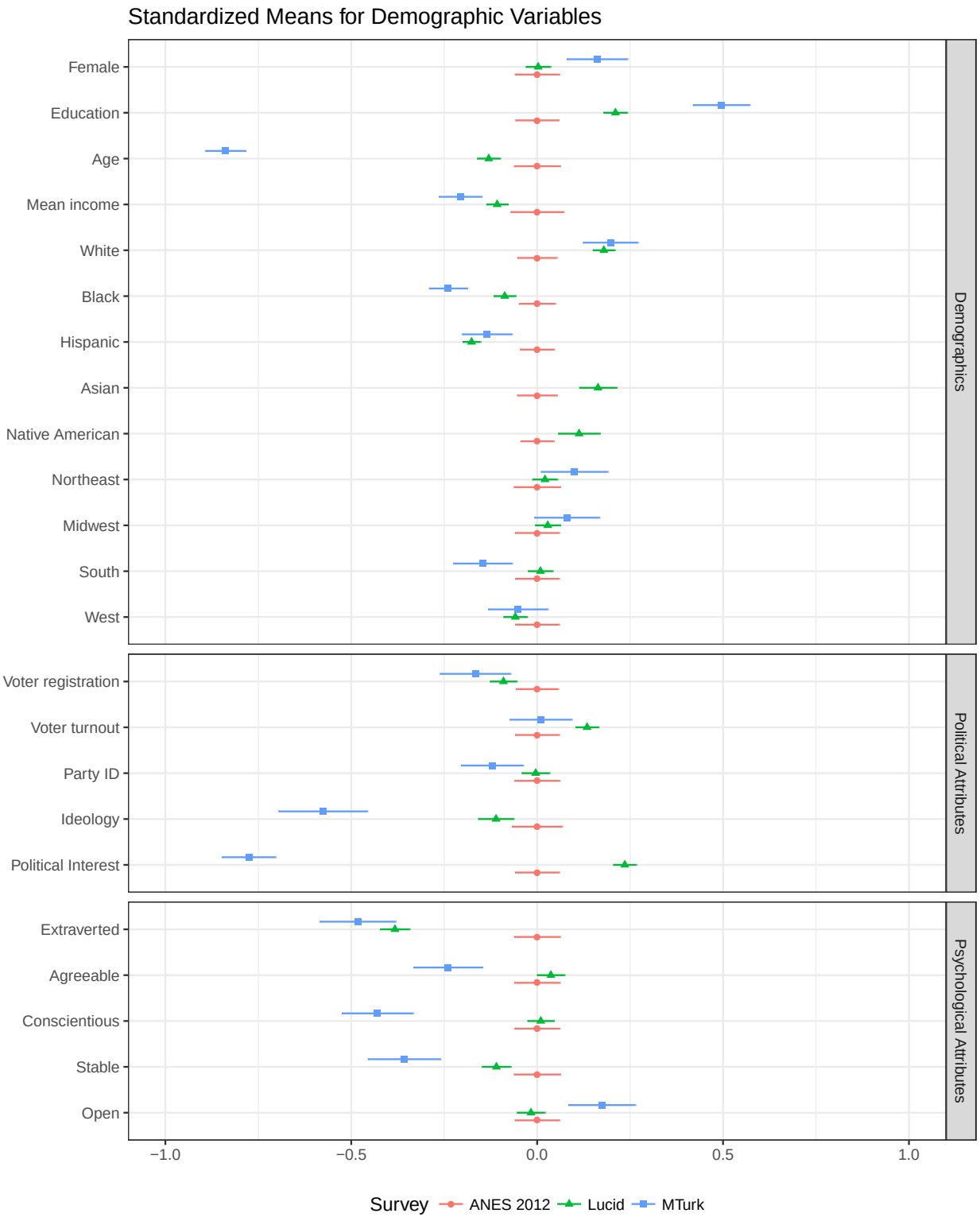

Figure 2: Summary of Experimental Comparisons Across Samples

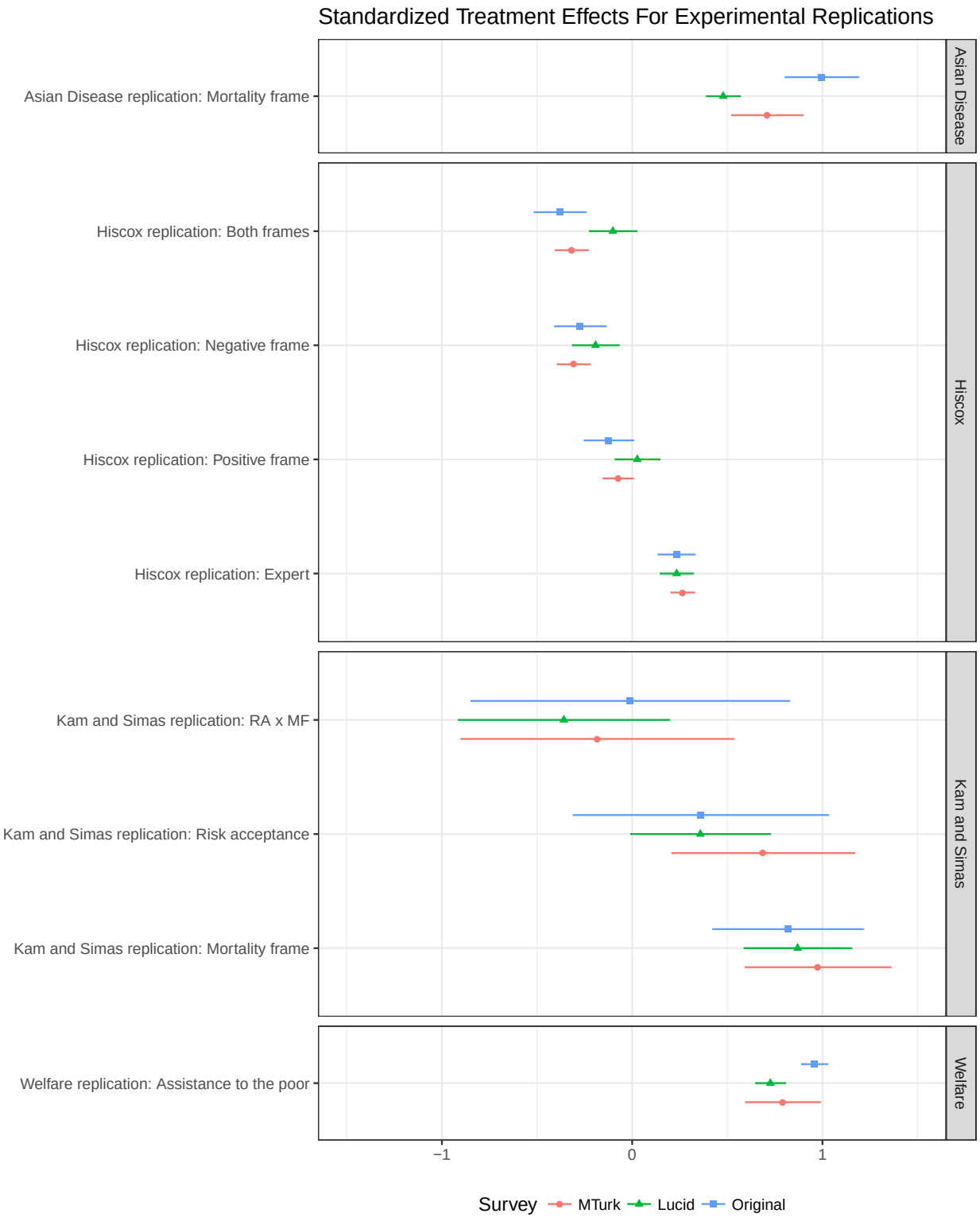

Table 9: Lucid sample survey behavior

	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.

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.

Finally, 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 to or different from Mechanical Turk. While we are reassured that on most dimensions, Lucid data appear to equal or outperform Mechanical Turk, 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. forthcoming. "Rumors, Truths, and Reality: A Study of Political Misinformation." *British Journal of Political Science* .

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.

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.

Cacioppo, John T and Richard E Petty. 1982. "The need for cognition." *Journal of personality and social psychology* 42(1):116.

Chandler, Jesse, Gabriele Paolacci, Eyal Peer, Pam Mueller and Kate Ratliff. in press. "Non-Naïve Participants Can Reduce Effect Sizes." *Psychological Science* .

Coppock, Alexander. 2016. Positive, Small, Homogeneous, and Durable: Political Persuasion in Response to Information PhD thesis Columbia University.

Flores, Alejandro and Alexander Coppock. 2016. "Do Bilinguals Respond More Favorably to Candidate Advertisements in English or in Spanish?".

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.

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.

Jarvis, W Blair G and Richard E Petty. 1996. "The Need to Evaluate." *Journal of Personality and Social Psychology* 70(1):172–194.

Kahan, Dan M. 2013. "Fooled Twice, Shame on Who? Problems with Mechanical Turk Study Samples, Part 2.".

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

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.

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.

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.