Representativeness versus Response Quality: Assessing Nine Opt-In Online Survey Samples

M.N. Stagnaro[1], J.N. Druckman[2], A.J. Berinsky[3], A.A. Arechar[4], R. Willer[5] and D.G. Rand[1,6]

[1] Sloan School of Management, Massachusetts institute of technology, United States; [2] Department of Political Science, University Rochester, United States; [3] Department of Political Science, Massachusetts institute of technology, United States; [4] Department of Economics, El Centro de Investigación y Docencia Económicas, Mexico; [5] Department of Sociology, Stanford University, United States; [6] Department of Brain Cognitive Science, Massachusetts Institute of Technology, United States.

* Corresponding author: mnstag@mit.edu

ABSTRACT Social scientists rely heavily on data collected from human participants via surveys or experiments. To obtain these data, many social scientists recruit participants from opt-in online panels that provide access to large numbers of people willing to complete tasks for modest compensation. In a large study (total *N*=13,053), we explore nine opt-in non-probability samples of American respondents drawn from panels widely used in social science research, comparing them on three dimensions: *response quality* (attention, effort, honesty, speeding, and attrition), *representativeness* (observable demographics, measured attitude typicality, and responding to experimental treatments), and *professionalism* (number of studies taken, frequency of taking studies, and modality of device on which the study is taken). We document substantial variation across these samples on each dimension. Most notably, we observe a clear tradeoff between sample representativeness and response quality (particularly regarding attention), such that samples with more attentive respondents tend to be less representative, and vice versa. Even so, we find that for some samples, this tension can be largely eliminated by adding modest attention filters to more representative samples. This and other insights enable us to provide a guide to help researchers decide which online opt-in sample is optimal given one's research question and constraints.

First version: Feb, 21, 2024
This version: Apr, 24, 2024

Introduction

Research in the social sciences relies heavily on human participants generating data via surveys and experiments. Researchers have increasingly used opt-in samples composed of people who joined a survey panel or completed a one-off survey while using the internet. These samples are often relatively easier to recruit, faster, and notably less expensive compared to traditional probability samples. But such samples may not be representative of any broader population they are used to study, threatening the generalizability of the inferences made. For researchers interested in the effects of experimental treatments and partial correlation relationships, some work suggests that opt-in non-probability samples may provide a reasonable alternative (Druckman and Kam 2011, Mullinix et al. 2015, Coppock et al. 2018, Druckman 2022, Jerit and Barabas, 2023).[1] Nonetheless, the response quality (e.g., attentiveness) and representativeness of opt-in on-line samples in social science remains an issue of central concern. Are some opt-in samples better than others? Are there tradeoffs, and if so, what are they and can they be mitigated?

Here, we address these questions by examining a variety of outcomes across nine different samples. Relative to previous studies evaluating non-probability samples (Horton et al., 2011; Berinsky et al. 2012, Mullinix et al. 2015, Coppock et al. 2018, Coppock and McClellan 2019, Kennedy et al. 2021, Douglas et al. 2023), we conducted a broad sample comparison where we (i) include a wider array of opt-in sample providers than most previous work, (ii) consider a larger array of sample characteristics, (iii) examine how screening respondents to increase participant attentiveness affects each sample's observable representativeness, and (iv) look at how efforts to make opt-in samples relatively more representative fail or succeed at bridging the gap between these types of samples and probability samples. We aim to provide social scientists interested in using opt-in samples a guide of which are best suited to their needs, based on their questions of interest and resource constraints.

There are several dimensions on which to assess opt-in samples. Here we have focused on three broad categories:

Response quality: We used the term "response quality" to refer to: how attentive participants are while taking the study, the level of honesty in their responses, the extent to which participants are willing to expend effort on the task presented, and respondents' willingness to take time and persist through the entire study.[2] These elements all are necessary to generate usable data for those who use these platforms (e.g., Hillygus and LaChapelle 2022); indeed, they are straightforwardly of value to researchers, especially with regard to experimental methods, ensuring participants receive key aspects of a given treatment (on the importance of attentiveness in experiments, even in light of mundane realism considerations, see Mutz 2011, Druckman 2022).

Representativeness: This category included the extent to which the demographic characteristics of the participants match the population of interest (e.g., the United States), the extent to which individuals of a given demographic subgroup hold the same preferences, beliefs, and attitude as those in the target population (attitude typicality), and the extent to which participants display similar patterns of experimental treatment effects compared to the target population. Representativeness is important for research that attempts to use a

[1] See Cornesse et al. (2020) and Jerit and Barabas (2023) for discussions of when non-probability opt-in samples are more or less fit for purpose.

[2] We avoid the term "data quality" since that envelopes more criteria, referring broadly to fitness for use (Bimer and Lyberg 2003).

sample to make claims about the wider population; e.g. a sample of older, rural republicans who all indicate being pro-choice, anti-gun and atheist would not be representative of a broader population of older, rural Republicans, despite sharing the same demographics. As we discuss below, assessing representativeness with opt-in samples comes with some inevitable caveats (e.g., we can only evaluate observable measures).

Professionalism: This category included the extent to which participants are infrequent versus frequent survey-takers (e.g. taking surveys in succession), the amount of experience participants have with this kind of survey, and the device on which they took the study. Aspects of professionalism can influence response quality (Kees et al. 2017), and/ or representativeness (Valentino et al. 2020). For example, frequent survey-takers who sit at a computer taking surveys in succession may not reflect the population of interest; they also may have been previously exposed to standard measures, manipulations, paradigms, or information, which could drive both domain specific knowledge (e.g. political knowledge), and task specific "immunity" (e.g. familiarity with some paradigms and measures, Rand et al, 2014). This could artificially enhance response quality. Alternatively, taking a survey on a mobile phone could lessen response quality given the screen size and likelihood of multi-tasking.[3]

In the current work, we compare a number of commonly used online samples in the social sciences across the above dimensions.

METHODS

Samples examined

We compare nine different samples of Americans. We aimed to recruit N=1,000 participants per sample, with the exception of Lucid, from which we aimed to recruit N=4,000 participants due to known high inattention rates on that platform[4]. In addition to eight professionally sourced samples, we also include one "in house" managed sample as an example of what a private social science lab can obtain. This resulted in a total collection of *N*=13,053 participants that started the study. Specific details for each sample can be found in SI-1.

Breakdown of the samples

We examine three different samples recruited from Amazon Mechanical Turk (Mturk) (Berinsky et al. 2012; Paolacci, Chandler & Ipeirotis, 2010; Horton, Rand & Zeckhauser, 2011): an unfiltered sample recruited directly from Mturk (*Open Mechanical Turk*), a sample filtered for quality using CloudResearch's approved list (*CloudResearch's Mturk Toolkit*), and a sample filtered for quality using an "in house" panel managed by the Polarization and Social Change lab at Stanford University as an example of what a private social science lab can obtain with their own Centralized Research Systems, Tools And Labor (*CRSTAL*). We examined two samples recruited from Prolific (Palan & Schitter, 2018): one sample balanced on gender (*Prolific*) and one sample using Prolific's "Representative Sample" option that provides a sample of Prolific users quota-matched to the national distribution on age, ethnicity, and gender (*Prolific NR*). We also

[3] It also might be the case that the presence of professional respondents is relevant given the purpose of specific a project (e.g., professional respondents tend to have less interest in politics and as such could make a sample recruited for a political survey more generally reflexive of political interest) (Hillygus et al. 2014).

[4] To ensure we had an ample number of individuals who passed the majority of attention checks for a direct comparison at that quality level, we needed to increase the Lucid collection notably. This was based on past experience with failed attention rates of between ⅓ - ⅔, given the length and content of this survey.

examined two samples recruited from CloudResearch's Connect panel: a pure convenience sample (*CloudResearch Connect*) and a sample from the Connect panel that is quota-matched to the national distribution on age, race/ ethnicity, and gender (*CloudResearch Connect NR*). Finally, we examined samples quota-matched to the national distribution on age, gender, ethnicity, and geographic region recruited from Lucid Marketplace (owned by Cint), an aggregator which recruits subjects from many different panels and sources (*Lucid*), and from Bovitz-Forthright (*Bovitz*), a survey firm who procures and maintains their own panel. Table 1 provides a summary of these different data collection efforts. We recognize that comparing quota versus non-quota, non-weighted samples introduce likely differences in sample composition. Our point is to assess these samples as they seem to be commonly used. We will later evaluate the impact of sample weights when estimating experimental treatment effects.

Table 1: Details of Samples[5]

Sample	N	Percent of All Data	Demographic Quotas
CloudResearch's Mturk Toolkit	1,088	8.34	None
CloudResearch Connect	1,046	8.01	None
Connect NR	1,028	7.88	Age, gender, race, ethnicity
Bovitz Forthright	1,115	8.54	Age, gender, race, ethnicity, geographic region
Lucid	4,505	34.51	Age, gender, race, ethnicity, geographic region
Open Mechanical Turk	1,103	8.45	None
Prolific	1,051	8.05	Gender
Prolific NR	1,040	7.97	Age, gender, ethnicity
Stanford CRSTAL Mturk panel	1,077	8.25	None

Recruitment procedure

Data were collected between June 2022 and November 2023. The overall study duration had a median time of 18.8 minutes for those who completed the full study, but, as we note below, varied notably across samples. In total, N=13,053 participants started the study. A total of N=105 individuals (not included in the total above) participated in the study more than once within the same sample; their first set of responses was used and all subsequent responses (for that sample) were excluded. An additional N=353 individuals participated in the study through multiple samples; all of their responses are included since that aligns with general usages where one does not know (unless it is asked) what other studies or samples participants have engaged with.

Design

All participants, from every platform, took the same survey. The survey included a series of demographic items; a number of attitude, belief and cognitive measures; several randomized experimental treatments; and ended with several items asking about participant's experience on

[5] Dates of data collection are provided in SI-1. Also, we considered additional data platforms (e.g., YouGov), however, we opted not to include them given: (i) an inability to independently control the data collection process (removing sample provider bias from the process), and (ii) it is our understanding that some providers, such as YouGov, conduct independent post-collection data cleaning for quality control as well as manually pruning samples to conform to quotas. While this is certainly useful for many purposes, it is notably different from the providers which we focus on here. Further, this introduces possible concerns for experimental studies (e.g., post-treatment conditioning).

the given platform and their experience taking this specific survey (for details on the survey flow, see SI-2). Throughout the survey, we asked multiple attention check and effort questions (described in detail below). Participants were not excluded for failing attention checks, as a key analysis we conduct below examines how responses change based on different approaches to filtering (removing those who miss attention checks). Thus, no respondent was dropped from our analyses (for any reason), unless otherwise noted. Question wording will be available at time of publication.

Compensation was set to the sample provider-specific standard for each sample for a study of this length. This rate varied across samples. This decision was based on, (i) the inability of some sample providers to modify their form of compensation (e.g., Lucid); and (ii) the variation between baseline compensation rates across samples meaning that we would not receive typical respondents/performance if we fixed all samples to a standard rate (e.g., standard Bovitz-Forthright rates would be unusually high if given to Mturk or Prolific participants, and standard Mturk rates would be notably low for Bovitz-Forthright participants). Therefore, though compensation amounts varied, the experience of participants from any given sample was typical for the sample of a given provider.

Variable constructs and coding

Response quality

In measuring response quality, we used the following constructs: (i) seven attention check items, distributed across the survey (with a total score of 0=missed all to 1=all correct); (ii) one reading-focused effort task, involving reading a paragraph of generic text followed by answering a single item question about that text, presented on the same page (0=failing to answer/ giving wrong answer and 1=correctly answering); (iii) one self-report item at end of the survey asking about honesty - if they had lied in any responses (0=reported lying, 0.5=reported not recalling[6], 1=reported honesty); (iv) a measure of speeding, identifying those who completed the study in less than 10 minutes[7] (0=speeding, 1=not speeding); and (v) a measure of attriting, defined as completing the initial demographics, entering the experimental section, but not finishing the study (0=attrited, 1=finished). We then create an aggregate response quality measure by averaging all of the above items (α=0.51 cross all samples, with notable variation discussed below).

Representativeness

As a benchmark against which to compare each sample's demographic and attitudinal characteristics, we used data from several recent probability samples (e.g. the 2021 General Social Survey (GSS), 2020 American National Election Studies (ANES) survey, Pew Research's Religious Landscape Study, etc.)[8]. For demographic representativeness, we examined age, income, education, race/ ethnicity, political party affiliation, and religious identity[9]. For each sample and each characteristic, our demographic item as asked in the study either matched, or

[6] The options for this item were 1-"Yes - at least once I was not honest", 2-"I don't remember", and 3-"No I was always honest". We gave those indicating the middle "I don't remember" answer a half score, as it likely indicated a "yes" but was not definitive. Though some readers may be surprised, previous work shows that participants who report lying exhibit other tendencies consistent with dishonesty are fairly open about lying on online surveys (Halevy, Shalvi & Verschuere, 2014).

[7] Given the median time was just over 18 minutes, below 10 minutes suggests a participant did not meaningfully engage with the full content of the study. This also loosely corresponds to the bottom 10% of speeds for finished studies, which was 10.16 minutes. Our speeding measure also roughly aligns with half the median completion time, a common threshold.

[8] See SI-3 for a full breakdown of which source supplied which characteristic.

[9] We did not include gender as it was a fairly easy target for all samples to match. With the exception of Open Mturk, the samples only varied by one or two percentage points on gender. See SI-4 for the distribution of gender.

was re-coded to be commensurate with, those of the probability sample[10]. We then calculated the absolute difference (in percentage points) in frequency between the sample and the benchmark at each possible level of the characteristic, and then summed these deviations across characteristic levels. For example, consider education; the following four levels make up U.S. educational attainment according to ANES 2020: "*Grade School/ Some High School''* (7.4%), "*High School Diploma*" (26.8%), "*Some College/ No Degree*" (29.2%), and "*College Degree/Post Grad''* (36.7%). These four levels make up 100% of the probability sample benchmark for education that we then contrast with each sample. Looking at the level of "*Grade School/ Some High School''*, we see the CRSTAL sample has 0.47% of participants with this level of education. We take the absolute difference with the probability sample benchmark to obtain a difference of 6.93. We sum this and the absolute difference for the other education levels ("*High School Diploma''*: AbsDiff= 17.45, "*Some College/ No Degree"*: AbsDiff= 8.27, "*College Degree/Post Grad''*: AbsDiff= 32.55) to arrive at a summed absolute difference of 65.2. We then computed the other summed absolute differences for age, income, race/ ethnicity, religious affiliation and political party, and averaged over all. This results in an aggregate *Representative characteristics* score (where higher values indicate a larger deviation from the benchmark, and thus a less representative sample).

For attitude/belief representativeness ("typicality"), we asked a series of items, examining death penalty position, gun control position, position on immigration, position on military spending, belief in God, reported political ideology, trust in people overall, reported voting behavior in 2016, and trust in the federal government. As described above for demographic representativeness, we calculate the absolute difference (in percentage points) in frequency between the sample and the benchmark for each response level, and then sum across all levels to calculate a score for each attitude. Then, as with the separate measure of representativeness on demographics above, we averaged across all attitude/belief measures to create an aggregate *Representative typicality* measure.[11]

Of course, in assessing demographic representativeness and attitude typicality, we could only look at representativeness of the *observable* variables that we measured (and our use of the term "representative" should be understood as such). The potential for unmeasured confounders is one reason why researchers turn to probability samples when the purpose of a study is to arrive at a precise population point estimate (MacInnis et al. 2018). Most of the work carried out on the platforms we explored are more interested in identifying relationships, often causal via experiments, between variables (Druckman 2022).

For this reason, we also looked at the ability for each sample to reproduce well known and robust treatment effects. Here, two experimental treatments were administered in random order. The first included a well-known survey experiment on welfare attitudes from the GSS. This involves showing participants a series of policies and randomly varied the wording of one experimental item, using the phrase: "*helping the poor*" versus "*welfare*" (Schuman & Presser 1981; Smith, 1987; Green & Kern, 2012). The second treatment involved a classic example from prospect theory, involving participants reading a description of a disease and a set of policies that contrasted concrete versus probabilistic outcomes and then choosing which of two policies they

[10] In some cases, for comparison purposes, we regrouped variable categories in these opt-in non-probability samples to facilitate comparability in comparison with the probability benchmarks.

[11] We recognize that samples will vary on specific items and these acute differences may not be apparent in our aggregate measure. We also provide results for specific items. Moreover, we recognize that any differences may reflect the distinct modes and/or timing of the probability surveys. While this might create a problem in assessing comparisons with the given benchmark sample, it is much less concerning given our focus in comparing across our opt-in samples (that were collected in a single mode and at similar times) and thus each sample faces the same variation from the probability benchmark.

would prefer (Tversky & Kahnaman, 1979; Kam & Simas, 2010). Both experiments have been widely tested, replicated and represent some of the more enduring and larger effects in their respective literatures (Berinsky, Huber, & Lenz, 2012). We also explore hypothesized interactions obtained via different demographics (e.g., political affiliation). As mentioned, in many cases, researchers who use the samples we study for the purpose of making an inference about a relationship. We thus have particular interest in variation in treatment effect estimates across the samples.

Professionalism

Participant professionalism was measured using a set of self report items. As a measure of experience, we asked how many similar studies that "asked about attitudes and beliefs" participants have completed through the current sample provider. We also asked participants what they were doing immediately before taking the study (e.g. completing a series of back to back studies versus working, socializing, shopping, etc.) to see whether they are professional versus incidental survey-takers. And lastly, we measure on which device the participant was using (mobile phone vs computer).

RESULTS

Response quality

Figure 1 shows average overall response quality (pooling across attention checks, effort task, self-report honesty, speeding, and attrition) as well as averages for each separate quality measure, across samples. Broadly, there are three groups of samples when it comes to measures of response quality. The highest quality group included the two Connect samples, the two Prolific samples, and the CRSTAL panel, which all had similarly high levels of overall response quality ($p > .05$ for all comparisons within this set; $p < .001$ for all other comparisons with the lower performing samples). Bovitz-Forthright and, to a lesser extent, CR Toolkit, occupied the middle quality range and were significantly different from each other ($p < .001$). Finally, quality was much lower on Lucid and Open Mturk. Lucid in particular scored the lowest on all sub components with the exception of honesty - which was lowest on Open Mturk, and was the lowest of all scores. Patterns for the individual quality measures were broadly similar, with the exception of Open Mturk.

How much can simple attentiveness filters at the outset of the study (to screen out bots and highly inattentive participants pre-treatment) help with response quality? We find that excluding participants who fail either of two (extremely easy captcha-style) attention checks at the study outset[12] increases overall response quality[13] (across samples, pre filtering *mean* = .771, *SD* = .26, *95%CI*[.77, .78]; post filtering *mean* = .822, *SD* = .21, *95%CI*[.82, .83]), but does not meaningfully change the rank ordering across all samples[14] (see blue diamonds in Figure 1).

[12] The front-end attention checks are maximally simple, designed to screen participants who are unambiguously not paying attention: "*Please write "twenty-five" using numbers*." and "*Help us keep track of who is paying attention. Please select "Somewhat disagree" from the options below.*"

[13] We used a modified quality score here that *removed* the first two attention checks for the overall score, as not doing so and subsetting on those who get the first two items correct would just mechanically produce a higher score.

[14] With the one exception of Prolific and CRSTAL swapping positions.

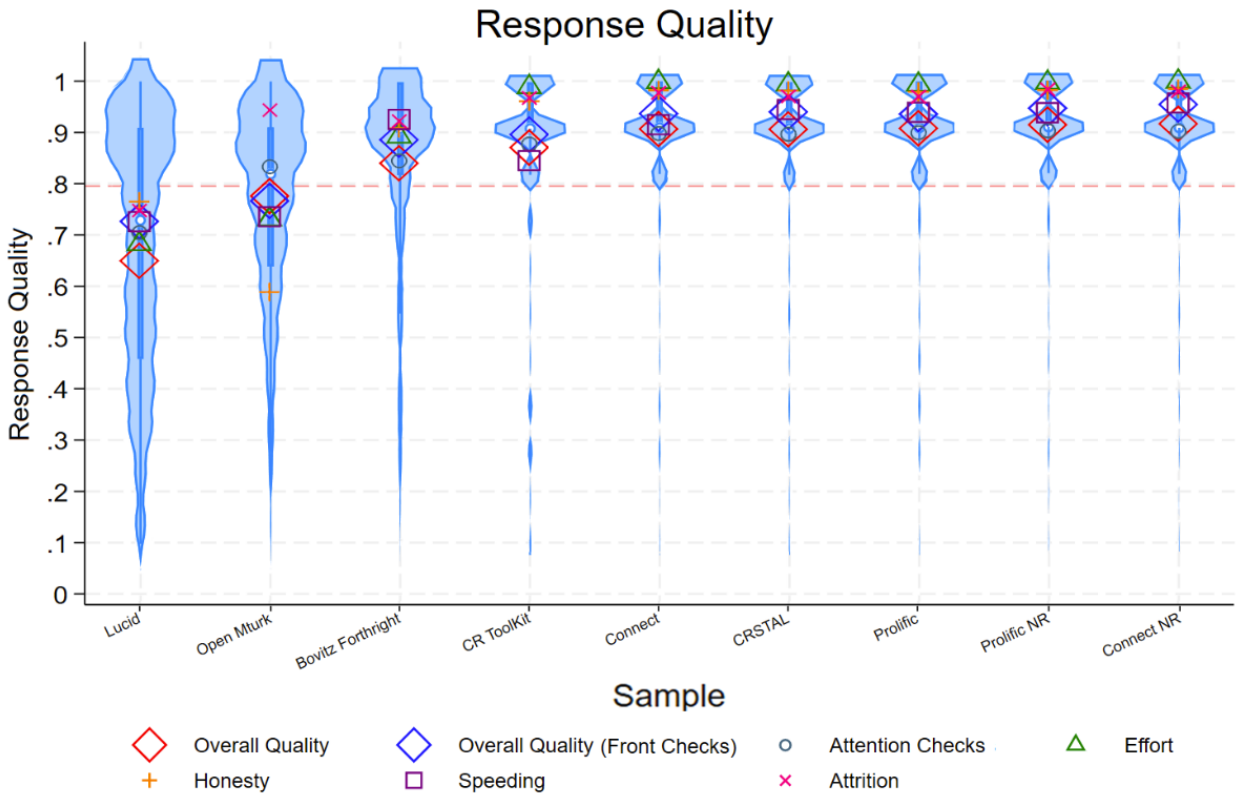


Figure 1. Variation in response quality across samples. *Violin plots depicting the distribution of overall quality by sample. Red dotted line indicates the overall average, across all samples. Red diamonds represent mean overall response quality, blue diamonds represent mean overall response quality after filtering on the first two attention checks. Other markers indicate the average for each subcomponent of response quality: aggregate of seven attention checks, reading effort, honesty, speeding, and attriting after demographics (all calculated with no front-end filtering).*

Seeing this improvement in response quality, one may ask if performance on the two front end attention checks identify a lower quality "type" of participant, as some have argued (Berinsky et al., 2024). To shed light on this question, we looked at Cronbach's alpha scores for the full set of seven attention check items[15]. Interestingly, there is comparatively little relationship in performance across attention checks for most of the six high data-quality samples (overall: *a*=.221, Connect: *a*=.26, Connect NR = .46, CloudResearch Toolkit: *a*=.19, CRSTAL panel: *a*=.28, standard Prolific: *a*=.05, Prolific NR: *a*=.05), but much more consistency across attention checks in the three samples with lower response quality (Bovitz-Forthright: *a* = .65, Lucid: *a* = .73, open Mturk: *a* = .58). This provides some evidence that weeding out inattentive participants in the latter three samples could remove an inattentive type of participants and thus yield substantial increases in quality (however, see the section below on filtering on attention and representativeness).[16]

It could be that these alphas are tracking some other signal (e.g., participants familiarity with attention checks in surveys). Therefore, we also looked at how well passing these front end checks predicts overall response quality[17]. If performance on attention checks is more about familiarity with how attention screeners work, there should be less of an association with overall response quality, as this aggregate includes a variety of items, including actual behavior (e.g.

[15] That is, we focus on the seven attention check items asked throughout the study.

[16] Interestingly, how fairly participants feel that they are compensated is not significantly associated with their response quality (see SI-3 for details).

[17] Response quality was calculated here after removing the front end attention items, to prevent auto-correlation.

speeding, attrition, effort expended, etc.). Whether participants passed both front end checks has meaningful predictive value for overall response quality on the three samples with lower response quality (Bovitz-Forthright: B = .484, R^2 = .23; Lucid: B = .615, R^2 = .38; open Mturk: B = .489, R^2 = .24), but much less so for the six samples with higher response quality (standard Connect: B = .288, R^2 = .08; Connect NR: B = .378, R^2 = .14; CloudResearch Toolkit: B = .4, R^2 = .16; CRSTAL panel: B = .367, R^2 = .14; standard Prolific: B = .234, R^2 = .05; Prolific NR: B = .305, R^2 = .09). Therefore, performance on front end attention checks predicts overall quality, and to a substantially greater externton samples with lower response quality. Clearly, early attention checks substantially help, particularly with samples that are otherwise low quality.

We next quantified the quality gains from implementing different numbers of attention checks. Table 2 shows the average overall response quality score for each sample when constraining to those who answered each number of attention check items correctly. These attention checks were all short, easy to answer, and had no trick/deceptive aspects; see SI-6 for wording (also see Thomas and Clifford 2017). We see that it is possible to achieve high response quality on all samples with a sufficiently high level of filtering (except, perhaps, open Mturk) - although this leads to a substantial loss of participants on the lower response quality samples (especially Lucid). This raises a question of whether attention filters that improve response quality might decrease observed representativeness. We next turn to this question.

Table 2. Means of overall response quality and percent of participants retained for each sample as a function of increasing attention filtering. *Scores are calculated using the following procedure: we first report raw quality scores, we then report quality scores filtering on those who answered the front two captcha-style attention checks correctly. Next, we report quality scores for those who correctly answered both of the captcha-style checks, plus any additional check correctly; then the front checks plus any two additional checks, and so on. Importantly, this comparison involves using a custom quality score at each level, such that it removes the attention check item being set to "correct" and averages across all variations when evaluating the overall score to prevent auto correlations. The second column of each grouping is the percentage of the original sample used, indicating the amount lost at each level of filtering.*

	Average data quality													
	Minimum number of attention checks passed (data quality & percentage of sample)													
	no filter		front 2		front 2 + 1		front 2 + 2		front 2 + 3		front 2 + 4		max filter	
Lucid	0.646	100%	0.723	69.8%	0.745	69.5%	0.737	65.8%	0.814	52.3%	0.783	39.4%	0.937	14.1%
OpenMturk	0.772	100%	0.763	87.0%	0.781	86.8%	0.784	84.4%	0.790	79.0%	0.737	64.4%	0.844	25.9%
Forthright	0.838	100%	0.883	88.7%	0.903	88.6%	0.903	87.0%	0.925	80.6%	0.852	69.9%	0.971	28.5%
CR Toolkit	0.868	100%	0.894	95.1%	0.915	95.1%	0.933	92.4%	0.950	88.5%	0.856	81.2%	0.965	35.0%
Connect	0.903	100%	0.933	96.7%	0.954	96.7%	0.962	95.5%	0.961	93.7%	0.861	86.4%	0.977	35.1%
CRSTAL	0.904	100%	0.938	96.6%	0.960	96.6%	0.962	95.9%	0.969	93.4%	0.869	87.0%	0.986	35.5%
Prolific	0.904	100%	0.932	96.5%	0.955	96.5%	0.958	95.4%	0.967	92.6%	0.874	85.3%	0.985	37.5%
Prolific_NR	0.909	100%	0.942	95.8%	0.964	95.8%	0.966	95.2%	0.960	93.9%	0.865	85.8%	0.977	36.3%
Connect_NR	0.913	100%	0.951	95.3%	0.974	95.3%	0.973	95.0%	0.970	93.7%	0.870	86.9%	0.989	36.2%
Total	0.851	100%	0.884	91.3%	0.906	91.2%	0.909	89.6%	0.923	85.3%	0.841	76.2%	0.959	31.6%

Representativeness

To compare the demographic composition of participants for each sample to a probability sample benchmark, we calculated the average absolute deviation (in percentage points) from that benchmark across six different demographics: age, income, education, race/ethnicity, political party affiliation and religious identity. Thus, lower scores indicate being *more* representative. We also look at each individual item disaggregated in the figure below. Figure 2 shows that Lucid and Bovitz-Forthright (which both use quota sampling) are substantially more representative than the pure convenience samples without quotas; and that the Connect NR and Prolific NR (which add limited quotas to Connect and Prolific, respectively) are somewhere in between, although closer

to the pure convenience samples than to Bovitz-Forthright and Lucid. (As these analyses yield only a single observation per sample, we do not conduct inferential statistics.) Lastly, we note that Open Mturk is substantially worse than all other samples.

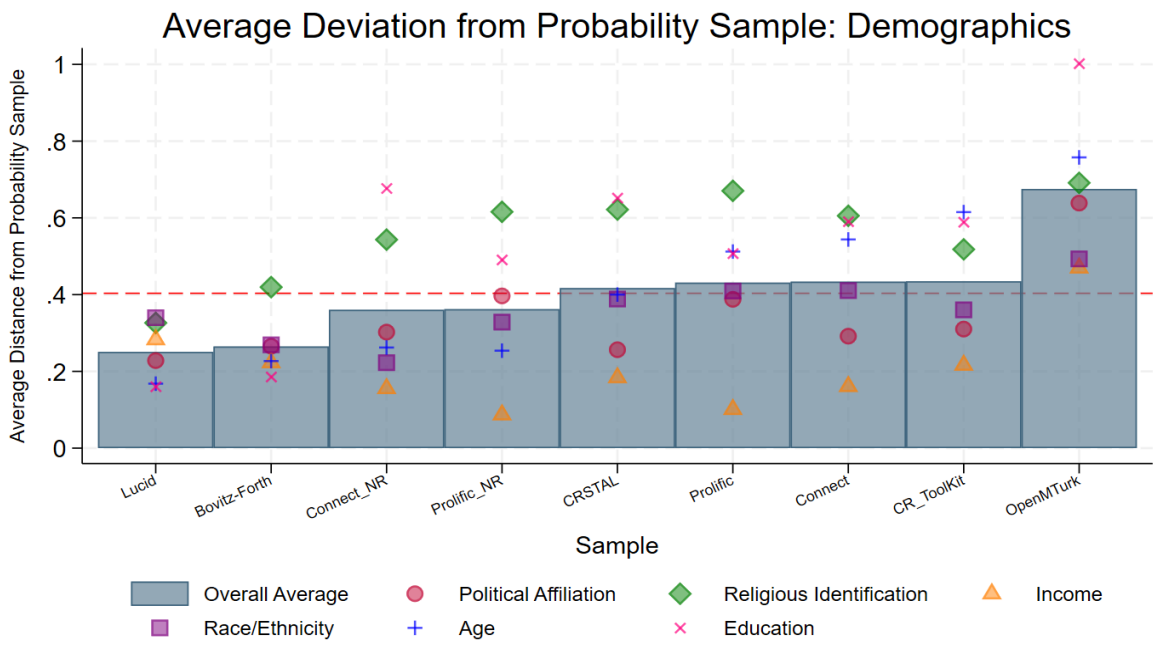


Figure 2. Deviation on demographics from probability sample benchmarks for each sample. *Bars indicate the average deviation from the probability sample benchmarks for each demographic across samples. Symbols depict each sub category: Political party affiliation, religious affiliation, income, race/ethnicity, age and education level. Red dotted line shows the average deviation collapsing across all samples. Y-axis shows average absolute deviations in percentage points.*

To compare how the samples vary on representativeness with regard to preferences, beliefs, and attitudes, for each sample we calculated the average absolute deviation (in percentage points) from the probability sample benchmark across nine different measures (death penalty position, gun control position, position on immigration, position on military spending, political ideology, belief in God, trust in people overall, reported voting behavior in 2016, and trust in federal government) (again also looking at the individual items disaggregated). Given the well-documented political divisions on these attitudes, we calculate these deviations separately by political affiliation (Democrats including independent Democrat learners, pure independents, Republicans including independent Republican leaners).

Figure 3 shows a similar pattern to what we observed above regarding demographic representativeness: Bovitz-forthright and Lucid are substantially more representative of preferences and beliefs than the pure convenience samples, and Connect NR is somewhere in between (although somewhat closer to the pure convenience samples). In contrast to our findings for demographic representativeness, however, Prolific NR offers little advantage beyond the pure convenience samples. Interestingly, the differences between samples are much more stark for Independents and Democrats than for Republicans. It is also striking that there are some topics where almost all samples notably deviate from the probability sample benchmark (e.g. belief in god for Democrats, voting in the 2016 election for independents, and gun control support for Republicans).

Average Deviation from Probability Samples: Typicality

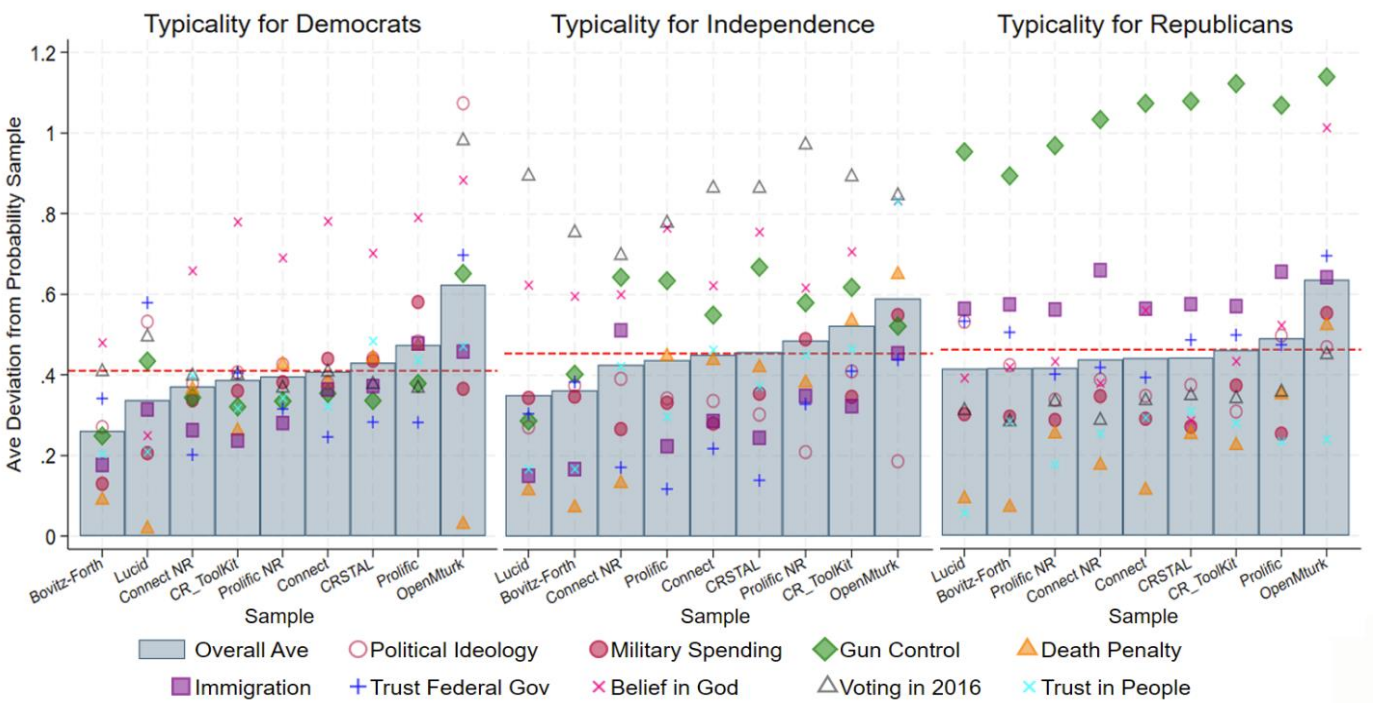


Figure 3. Deviation on typicality from a probability sample benchmark for each sample among Democrats, Independents, and Republicans. *Bars indicate the overall comparison, while each symbol disaggregates this score and shows the deviation for: death penalty position, belief in God, gun control position, position on immigration, position on military spending, political ideology, trust in people overall, reported voting behavior in 2016, and trust in federal government. Red dotted lines indicate average absolute deviation collapsing across all samples. From left to right, plots represent scores for Republican plus Republican learners, pure independents, and Democrats plus Democrat learners.*

Comparing these results with results presented in the previous section, there appears to be a notable negative relationship across samples between quality and representativeness. Compared to the other samples, both Lucid and Bovitz-Forthright scored notably worse on quality, but are notably closer to the probability benchmarks on representativeness. The one obvious exception to this pattern is Open Mturk, which both scored second worst on overall quality and was consistently the farthest (by a notable magnitude) from probability benchmarks.

We next examined how increasing response quality by filtering out inattentive participants affects representativeness. Figure 4 shows that, perhaps surprisingly, there is little change in average demographic representativeness across all but the most extreme levels of attention filtering. The pattern is similar when looking at the individual demographic variables, except for Lucid, where the lack of change in average demographic representativeness hides decreases in representativeness for some characteristics (e.g. education, income) and increases in representativeness for others (e.g. religiosity, age); see SI-8. This is also mostly true for typicality; here we see more variation with more extreme filtering, and again, Lucid shows notable swings in representation (see SI-9). Thus, including filters seems to enhance response quality without undermining representativeness, as long as the filtering is not too extreme (with filtering impacting representativeness on Lucid as a lower level of extremity than the other samples). This is a key insight given it provides a mechanism by which to address the apparent representativeness-attentiveness tension[18].

[18] See SI-8 for representativeness broken down by the levels of the other response quality constructs.

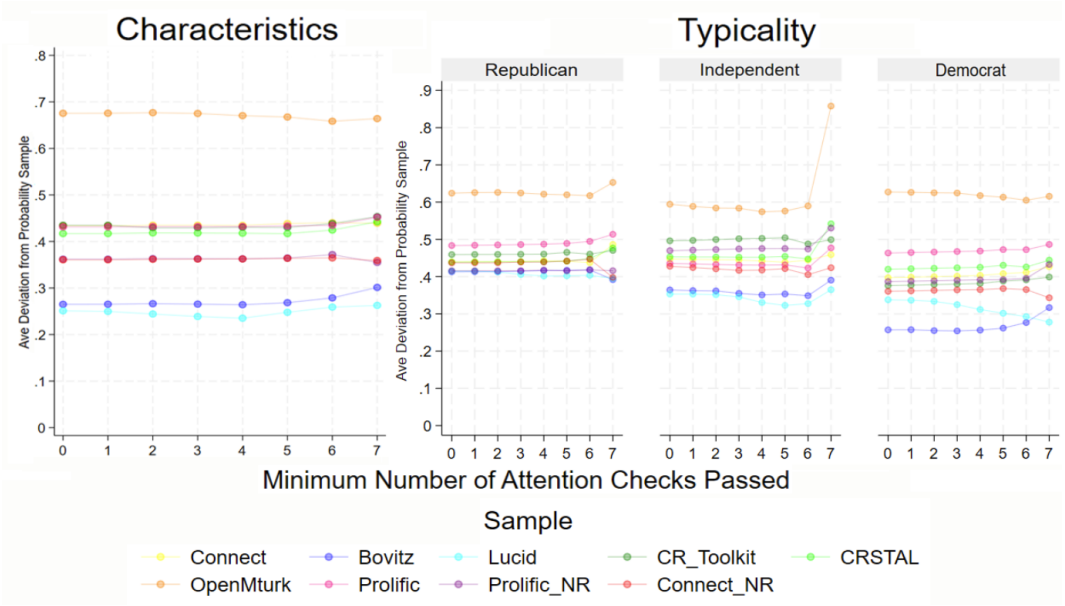


Figure 4. Response quality can be improved by filtering on attentiveness without (mostly) compromising representativeness. *The three plots on the left depict average deviation in typically for each sample, across levels of attention filtering. Attention levels represent filtering on the minimum number of attention checks correct, such that "0" is the full sample and "7" are those who passed all 7 attention checks. Note that these averages represent composers of several sub-categories which themselves are fairly stable across filtering, with the notable exception of Lucid, see SI-8 & SI-9.*

Finally, we consider how treatment effect heterogeneity varies across samples. Some previous work comparing probability versus opt-in non-probability samples has shown little variation in the ability to recover main effects (Mullinix et al. 2015, Coppock et al. 2018, Coppock 2019, 2022). However, others have argued that some failed replications are caused by the use of non-representative samples (Kahan & Peters, 2017; Kahan, 2015), with some specifically noting heterogeneity as the key concern (Tipton, Yeager, Iachan & Schneider, 2019). This finding could be a result of variation in treatment effect size across different subgroups (e.g. smaller effect sizes in younger participants, shrinking the average treatment effect if they made up a sizable part of the sample); or could be due to non-typical preferences and attitudes of particular groups (e.g. a sample of mostly fiscally conservative but socially liberal Republicans producing a different average treatment effect compared to more typical Republicans)[19]. Here, we investigate how treatment effect heterogeneity varies across samples by examining variation in two well established treatment effects and associated interactions with demographics; specifically, (i) contrasting the effect of framing a policy as "helping the poor" versus "increasing welfare", and (ii) participants' preference for concrete gains over probabilistic losses[20]. In these analyses, we do not filter on any attention checks.

"Helping the poor" vs "Those on welfare"

We first look to replicate an often replicated experiment on social entitlement spending (Schuman & Presser, 1981; Green & Kern 2012). Previous work has found greater public support when a policy is described as "helping the poor" compared to when described as "increasing welfare". Furthermore, this effect has been found to be much stronger for Republicans compared to Democrats.

[19] For a broader conversation see (Berinsky, Huber & Lenz, 2012).

[20] We also ran a third experiment from Malhotra and Popp (2012); however, we did not replicate the original effects on any platform and thus leave the results to the supplement.

We present participants with several policy topics (e.g. "*halting rising crime rates", "Social Security", etc.*) while asking "*Are we spending too much, too little, or about the right amount on this topic*". Participants then respond by selecting one of these three options. The final item asks either about "*welfare"* or "*helping the poor*" as the critical item.

Using ordinal logistic regression, we replicate the expected main effect collapsing across samples (*b*= -.883, *p*< .001, *95%CI* [-.956, -.811]). Though effect sizes vary in magnitude and precision, we see the effect also replicates across each sample individually, with the largest effect being Bovitz-Forthright (*b*= -1.17, *p*< .001), and the smallest being Open Mutrk (*b*= -0.2, *p*= .094) (see SI-10). Adding partisanship and the interaction between treatment and partisanship to the model also replicated the expected interaction (*p* < .001) (Green and Kern 2012), such that Republicans showed a much stronger framing effect (*b*= -1.18, *p*< .001) compared to Democrats (*b*= -.819, *p*< .001). Looking across samples, however, the interaction with partisanship was largest, and (at least nearly) significant, in the two more representative samples (Lucid: *b*= -.749, *p*< .001, Bovitz-Forthright: *b*= -.588, *p*=.051), versus smaller and not statistically significant in the other samples - and even opposite signed on Prolific; see Figure 5a.

Beyond looking at moderation by political affiliation, we also look at several other notable demographic groupings (age, ethnicity, and gender). One of the more pressing concerns for samples of varying representativeness is the ability to observe heterogeneous effects. Here we find, consistent with Green and Kern (2012), the more representative samples reveal moderation by age (above vs below 40; Figure 5c). We also find evidence of moderation by race (white vs non-white; Figure 5b), concentrated in the more representative samples.[21] For the non-representative samples, the range of effects show no consistency, and in some cases show significant effects for politics, age, or race. (We find no evidence of gender moderation in any sample (Figure 5d).

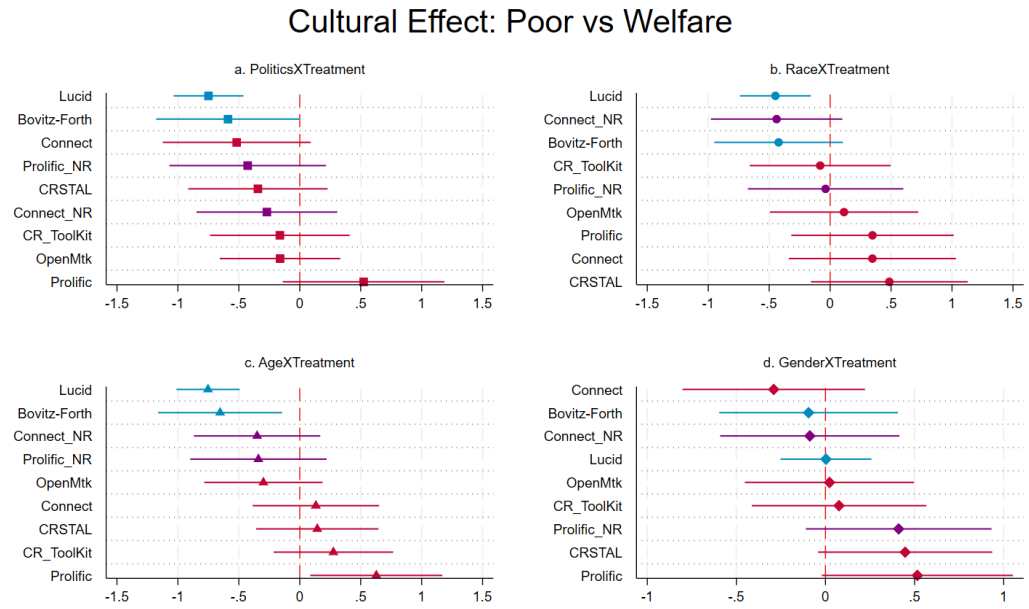


Figure 5. More representative samples show different patterns of individual difference moderation for the effect of welfare framing on policy support. *For each sample, we plot the interaction between treatment (framing the policy as increasing welfare versus help for the poor) and participant (A) party affiliation (Rep./ Dem.) with negative scores indicating more Democratic support, (B) race (white/ not white) with negative scores indicating more non-white support, and (C) age variable (above/ below 40 years) with negative scores indicating younger support; as well as (D) gender (female/ male) with negative scores indicating more female support. Blue indicates the two most representative samples*

[21] Green and Kern (2012) do not look at race as a moderator.

(Bovitz and Lucid), purple indicates the two somewhat representative samples (Prolific_NR and Connect_NR), and red depicts the least representative samples.

Preference for concrete wins over probabilistic losses

We also replicated the classic framing effect from Prospect Theory (Tversky & Kahnaman 1979), whereby participants prefer a certain outcome over a probabilistic outcome of equal expected value when the outcomes are framed as gains, but prefer the probabilistic option over the certain option when the outcomes are framed as losses. This effect has been widely replicated (e.g., Kuhberger 1998, Druckman 2001, Berinsky et al. 2012) and shown to operate in policy-relevant contexts (Kam & Simas, 2010). In our experiment, participants choose between two interventions: intervention A, where 200 people will be saved/400 people will die (concrete outcome) vs intervention B, where there is a 1⁄3 probability that 600 people will be saved/0 people die (probabilistic outcome), and there is a ⅔ probability that 0 people will be saved/600 people will die (see SI-11 for details).

Using a logistic regression, we predict the probability of selecting the probabilistic option over the concrete option as a function of frame (loss vs gain). We successfully replicate previous work, finding a significant effect of frame overall (b= .988, p< .001, 95%CI[.88, 1.1]), and in each sample, with the largest effect being Prolific: b=1.46, p< .001, and the smallest effect being open Mturk: b= .627, p= .001 (see SI-12).

Turning to heterogeneous treatment effects, previous work has identified an interaction between frame and risk preferences (Kam & Simas, 2010) (more details in the SI-13). In line with this work, we find a (barely) significant interaction between frame and risk preferences when predicting choice if the probabilistic option overall (p=.04), such that those below the median level of risk show a larger framing effect (b= 1.12, p< .001, *95%CI*[.957, 1.28]), compared to those above the median level of risk (b= .907, p< .001, *95%CI*[.756, 1.06]). The interaction effect is not statistically significant in any of the individual samples, and shows some evidence of variation across samples (Heterogeneity chi-squared between p = 0.003), though no sample is significant SI-13.

In contrast to the welfare framing experiment, Figure 6 reveals little evidence of any moderation between treatment and any of the demographic variables, and no evidence of significant variation across samples (heterogeneity chi-squared between p = 0.345 and p = 0.947). Also, unlike the other experiment, there is no obvious pattern whereby the more representative samples agree on an effect size or direction.

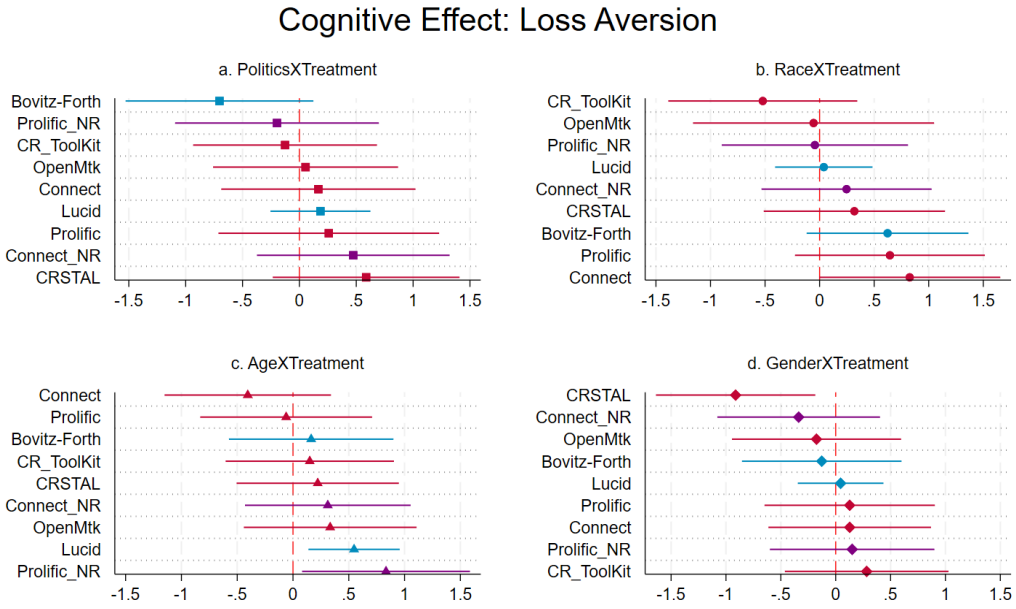


Figure 6. No clear relationship with representative samples on patterns of individual difference moderation for the effect of loss frame on policy support. *a. plots the interaction coefficient between treatment and party affiliation (rep/ dem), across all eight panels; b. plots the interaction coefficient between treatment and dichotomous race (white/ not white) variable; c. plots the interaction coefficient between treatment and dichotomous age variable (above/ below 40); d. gender showed no interaction but the main effect on outcome followed the same pattern. Blue indicates the two high representative samples consistently patterning together (Bovitz and Lucid), purple indicates the mid representative samples performing midway between both groups (Prolific_NR and Connect_NR) for typicality and demographics, and red depicts all non-representative samples.*

One reason we may have seen the interactions with sample representativeness in the first experiment (helping the poor vs increasing welfare), but not the second (loss aversion), may be due to the specific nature of the experimental content (Clifford et al. 2015). The former experiment involved social/political content; whereas the latter experiment has more of a cognitive focus with probabilistic framing. With this in mind, we present relationships documented in probability samples that are only recovered in the more representative samples: such as the notably diminished relationship between political ideology and party affiliation for Black Americans as compared to the strong correlation for White Americans (Jefferson, 2020) (Figure 7a); and the notable male bias in support for Donald Trump across the U.S. electorate (Chaturvedi, 2016) (Figure 7b).

a. b.

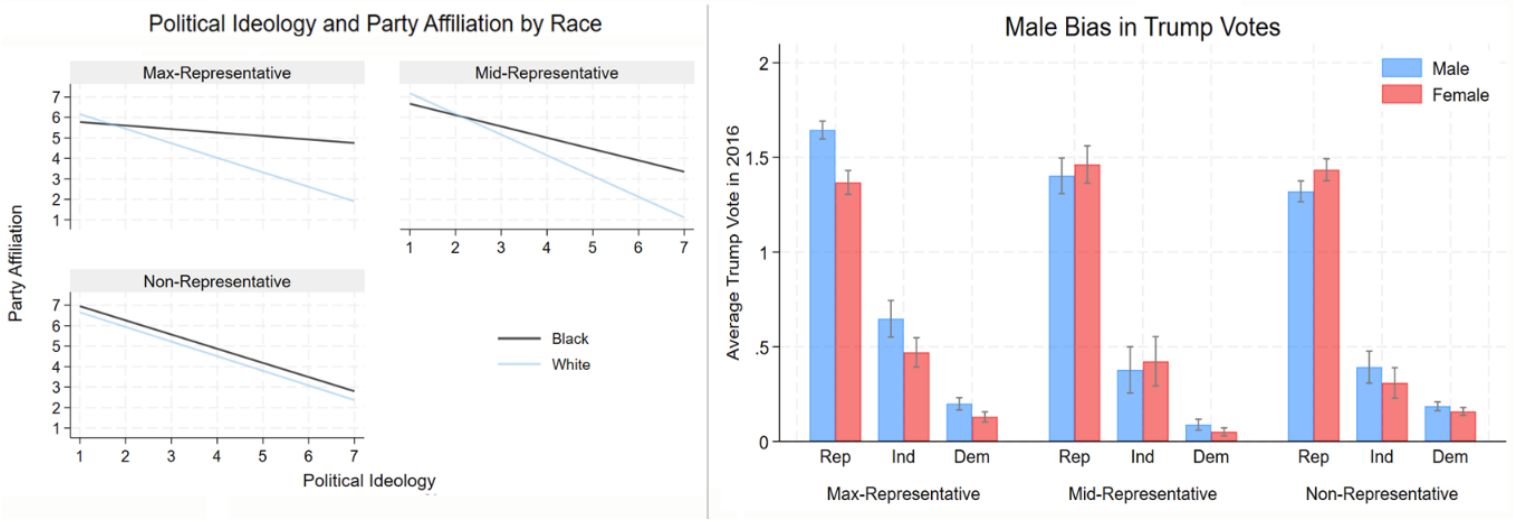


Figure 7. Additional comparisons of sample representativeness and capturing social/political content. *a. plots black participants (black line) showing little relationship between political ideology (x axis) and political party affiliation (y axis) for both the Max-representative (Lucid and Bovitz-Forthright), and to a lesser degree the Mid-representative samples (Connect_NR and Prolific_NR). As were both black and white (blue line) participants show a strong relationship between political ideology and political party affiliation on the non-representative samples (Open MTurk, CR_ToolKit, Connect, Prolific and CRSTAL). b. plots the relationship between participant gender and voting for Donald Trump. The max-representative samples capture this bias, as where the Mid and non-representative samples do not.*

Given that we only present two experimental treatments above, along with these additional analyses replicating other relationships of interest, we also ran an exhaustive analysis in the SI-14 using six key demographics (race, age, edu, gender, income and political affiliation) to predict the 22 main attitudes, beliefs, and cognitive measure we collected. This provides the reader with an exhaustive and unbiased comparison for any and all comparisons (on observed variables) one

may be interested in[22]. Across all relationships, we see some discrepancies between the more versus less representative samples (SI-14).

Sample Weights

One solution to the lack of sample representativeness is to apply post-stratification weights. Weights mechanically improve representation on the variables weighted; but, as a general matter, have limitations in terms of ensuring agreement with descriptive benchmarks. MacInnis et al. (2018) offer an extensive comparison of non-probability and probability samples, concluding that "post stratification weights …only sometimes improved the accuracy of the nonprobability samples. Furthermore, weighting did not eliminate the superiority of the probability sample surveys over the nonprobability sample surveys…" (p. 726). As mentioned though, the primary usage of most of the samples we study are not for precise population estimates but rather for studying relationships between variables, often with experiments. Some evidence suggests that many experimental effects are homogenous (Coppock et al. 2018), which would mean that the type of sample employed and/or the use of weights would not influence the estimated treatment effect (Druckman 2022; also see Miratrix et al. 2018). Our loss aversion experiment aligns with this view. Yet, the helping the poor vs increasing welfare experiment includes heterogeneity, and as such, we found inconsistent main and moderated effects across samples. In SI-21, we assess the impact of adding weights to each sample for this experiment. We find that the weights improve consistency for the main effect, such that every sample replicates the main effect after weighting (most notably affecting Open Mturk). Critically, however, there continues to be inconsistent moderated effects even after weighting, such that the weights do not successfully recover treatment effect heterogeneity. This is the case even when we combine all the less representative samples in an attempt to counteract the loss of power that comes from weighting. Therefore, we find no evidence that weithing closes the gap between the higher and lower representative samples in terms of heterogeneous effects.

Professionalism

Finally, we examine variation across samples in levels of professionalism/prior experience of the participants. Specifically, we examine the following variables: taking the study on a mobile phone versus laptop/computer; what activities participants were engaged in just prior to taking the study; past experience taking studies of this nature.

Starting with the device used to take the survey, Lucid and Bovitz-Forthright have a notably larger percentage of participants taking the study on mobile (70.6% and 69%, respectively) compared to the other samples: Connect (9.6%), Connect_NR (9.6%) CR_Toolkit (4.8%), CRSTAL panel (7.6%), Open mturk (0.5%), Prolific (20.5%), and Prolific_NR (18.9%). A regression predicting a user's response quality index with a mobile device dummy (and including sample dummies) finds that response quality is significantly lower on mobile (B= -.139, p< .001, *95%CI*[-.077, -.058]). Breaking each sample out, we see this difference is relatively stable across samples; however, it is notably larger for those samples with lower response quality (see SI-15). This difference suggests an interesting potential mechanism that could contribute to the observed differences in representativeness and response quality. The high rate of mobile phone use potentially extends the participant provider's reach, allowing access to more participants/a wider array of participants.

However, mobile phones are also a clearly suboptimal device for paying attention. This goes to the tradeoff between representativeness and response quality. Indeed, even filtering on

[22] Though, by design, there are no hypotheses being tested here, we still encourage the use of caution in interpreting any one result, given the large number of tests being run here.

attention checks is insufficient to remove the relationship between using a mobile device and response quality: regressing the response quality index on dummy variables for the platforms and mobile use, among those who passed the front end attention checks yields a significant relationship (*B*= -.097, *p*< .001, *95%CI*[-.04, -.03]) (which is slightly improved relative to those who did not pass the front-end attention checks (*B*= -.139, *p*< .001, *95%CI*[-.08, -.06]), but still reflects a decline in response quality relative to no mobile use).

Turning to what participants were engaged in prior to taking the study, participants could select as many options as they wanted out of a set including “completing other studies”, “working on the computer”, and a variety of other non-study related activities. Table 2 shows the breakdown of responses by sample. Overall, participants on the more representative samples are less likely to be working on the computer or completing other studies prior to the current task, suggesting a lower percentage of “professional” study participants in those samples (i.e. participants who were completing our study incidentally to other activities, rather than completing studies as their main activity).

Table 2. Percentage of users in each sample engaging in a given activity before taking the study.

	CR Toolkit	CRSTAL	Prolific	Prolific NR	Lucid	Open MTurk	Connect NR	Connect	Bovitz-Forthright	total
Completing other survey	46.49	45.69	38.08	32.64	31.49	30.71	25.75	25.53	12.9	32.04
Working on computer	32.13	30.69	29.66	27.96	16.07	46.86	36.82	43.84	20	28.67
Doing stuff on phone	3.51	5.88	8.32	11.54	15.19	2.59	9.36	8.71	29.5	11.38
Playing a game	2.21	1.47	3.31	2.99	12.61	4.79	4.23	2.7	8.2	6.18
Shopping	0.7	1.08	1	1.59	3.4	1.6	1.91	0.8	2.3	1.93
Socializing	1.81	1.76	2.3	3.78	3.46	1.4	2.92	2	6.4	2.98
House Work	1.41	1.47	1.3	1.59	0.88	0.1	1.91	1.4	1.1	1.17
Misc Work	0.3	0.59	1.1	0.7	0.91	0	0.91	1	3.1	0.95
Eating	1.41	2.84	2.71	2.29	1.34	0.2	2.62	3.1	2.3	1.95
Resting	1.2	1.96	2.61	3.58	3.23	0.5	4.83	1.7	4.8	2.81
Other	3.11	4.02	4.71	6.87	3.27	1.2	4.23	4	5.5	3.95
Mix of above	5.72	2.55	4.91	4.48	8.13	10.07	4.53	5.21	3.9	5.99
Total	100 (n=996)	100 (n=1,020)	100 (n=998)	100 (n=1,005)	100 (n=3,061)	100 (n=1,003)	100 (n=994)	100 (n=999)	100 (n=1,000)	100 (n=11,076)

Next, we consider the self-reported number of previous similar studies completed. Specifically, participants were asked the number of studies they have completed that contained content about beliefs, attitudes, or academic content similar to the content in this survey. We see wide variation in past experience across samples, with Bovitz-Forthright (median = 5) and Lucid (median = 4) on the low extreme, and the CRSTAL panel (median = 300) and CR-Toolkit (median = 200) on the high extreme (Figure 9). Regressing response quality against the number of studies completed (log10-transformed), including sample dummies, shows a significant positive relationship (*B*= .124, *p*< .001).

In sum, the inclusion of mobile access is associated with improved representation. Yet, it also is associated with lower response quality, which cannot be fully addressed with attention checks. These more representative samples also tend to have fewer repeat survey takers -- which depending on the purpose of the study can be a positive or a negative (see, e.g., Hillygus et al. 2014).

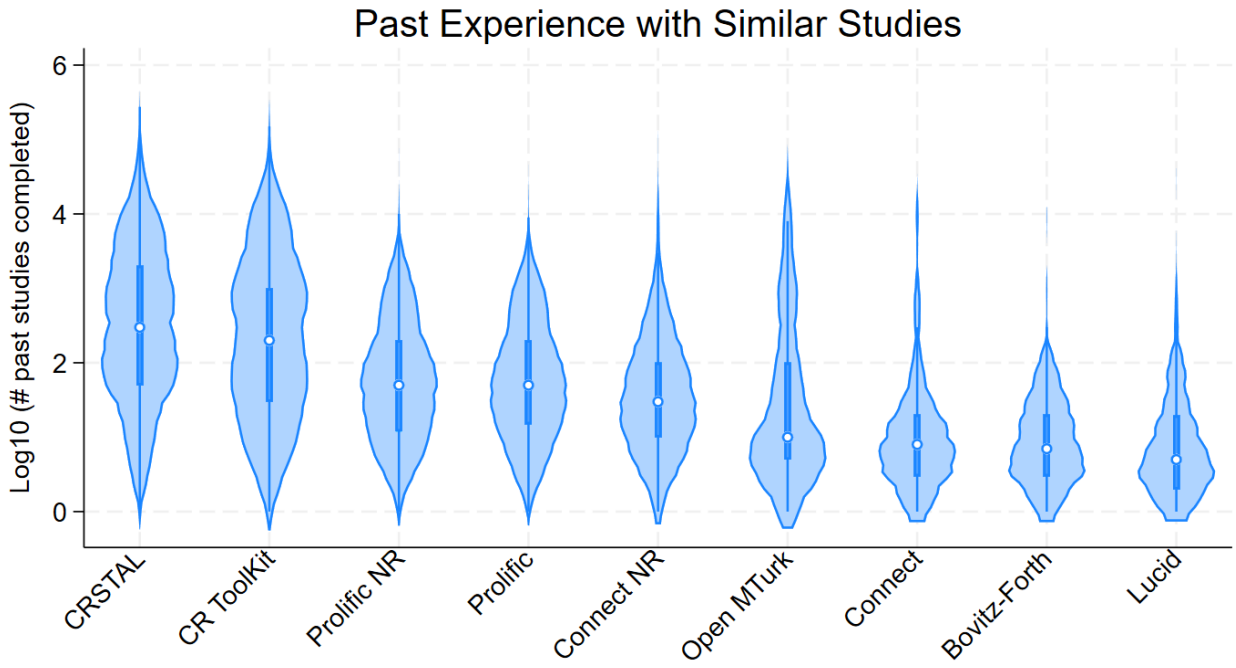


Figure 9. Level of past experience with similar studies varies widely across samples. Shown are violin plots of log10-transformed number of past similar studies participants reported having completed, by sample.

DISCUSSION

We have compared response quality (including attention, speeding, willingness to work, attrition and honesty), participant representativeness (including demographic makeup, attitude typicality, and the ability to recover treatment effects and key interactions), and participant professionalism (the use of cellphones, total past experience with academic studies, and taking other studies immediately prior to taking the current study) across nine different online samples. We find substantial variation across samples on all dimensions.

We find, in the samples we explored, a negative relationship between overall response quality and participant representativeness (r = -.66[23], see SI-18), as well as with non-professionalism (r = -.26). This illustrates a clear trade-off for researchers trying to choose which sample to use and which individuals to flag as inattentive (also see Clifford and Jerit 2015). The samples that were closest to representative probability samples on demographic and attitudinal representativeness (Lucid and Bovitz-Forthright) showed notably lower levels of response quality than the less representative samples. The quota-matched panels offered by CloudResearch and Prolific retained high response quality while being somewhat more representative than their non-quota-matched counterparts, but still lagged substantially behind Lucid and Bovitz-Forthright (especially on demographic representativeness). Conversely, filtering out inattentive participants improved response quality in the more representative samples. For Bovitz-Forthright, this culling was achieved by merely filtering on two trivial attention checks at the study outset, without compromising representativeness or losing many participants. For Lucid, however, imposing enough filters to reach response quality parity with the other samples caused a decrease in representativeness (as well as a substantial loss of participants).

We also find that applying weights to the less representative samples did *not* lead to consistent improvements in agreement with the more representative samples in terms of heterogeneous treatment effects (see SI section SI-21). The lesson is that if one is confident, a

[23] Removing OpenMturk which is an extreme outlier in the negative direction for both dimensions, see SI-18.

priori, of homogenous effects, then the choice of sample will matter little, at least in terms of representativeness. If there is the possibility of heterogeneous effects, weights are an imperfect solution, particularly given that they reduce power in a situation where power (to identify heterogeneous treatment effects) is already difficult to come by (Gelman 2018). Just how often heterogeneous effects occur in social science experiments is debated. Some work offers a compelling case for homogenous effects, at least regarding the provision of information and considering standard demographic variables (e.g., age, gender, education) (Coppock et al. 2018). That said, there is reason to expect more heterogeneity once accounting for distinct types of treatments and moderators (e.g., those more specific to the topic being studied such as issue importance, or motivational/skill/experiential based variables) (Druckman 2022, Krefeld-Schwalb et al. 2024). Generally, with regard to social, cultural, and political behavior questions (e.g., Bryan et al. 2021, p. 980) call for "the recognition that most treatment effects are heterogeneous."

More generally, the common representativeness and response quality trade-off demonstrates the importance of researchers thinking clearly about what sample characteristics matter most for any given study and outcome(s) of interest, and thus which sample is optimal for them to use. High response quality may be particularly important for studies with complicated designs, or treatments/measures that require participants to invest substantial effort or attention. Representativeness may be particularly important for studies where individual difference correlations or heterogeneous treatment effects are of interest, especially if the study involves social or political content (where pronounced individual differences are more likely, relative to content that is largely cognitive). Note that the focus of this project was to inform researchers about the characteristics of samples of individuals intended to be participants in a given research study, such as experiments. There are other reasons to collect a group of individuals from many sample providers for some form of online labor, such as image labeling or fact checking. Though this paper may still be of use to such a reader, we do not make any claims about these sorts of collections.

Finally, in addition to practical relevance, the observed trade-off pattern raises interesting theoretical questions: Why do we find that more experienced participants are less representative and produce higher quality data? How much of this association is caused by some type of learning process, versus a process of selection whereby people who are less attentive, and are from particular demographic and attitudinal subgroups, drop out of the sample after completing a relatively small number of studies (or simply complete studies less regularly)? If the issue is driven by the underrepresentation of some demographic groups, then adding a greater number of quotas to a collection could help close this gap. Indeed, this approach has been the strategy of many survey firms who use a large number of joint or marginal quotas, as well as quotas based on demographics and attitudes/ behaviors (e.g. voting turnout).[24] However, if the loss in representativeness is driven by some kind of learning or "deeper" change in the participant base, the answer may be more about minimizing exposure and keeping participants less experienced. While our weighting results bring the underrepresentation possibility into question, future research is needed to distinguish between the possibilities.

Additional platform features

Beyond response quality, representativeness, and professionalism, the samples we examined varied on other features that may be of interest. One key limitation of Lucid relative to the other samples is the inability to provide additional payments to participants based on their

[24] And in fact, as a consequence of this paper, some sample providers are now adding more demographic quotas to their default collection option in a direct effort to try and improve representativeness (e.g. Connect NR).

responses in the survey (thus making Lucid unsuitable for various experimental designs, such as incentivized economic games). Relatedly, while participants on all other platforms were compensated using monetary payments, participants on Lucid were compensated with a large array of currencies (e.g. money, coupons, video game coin, and even nothing), with compensation type not being under the investigator's control. Another key difference across samples is the feasibility of collecting multiple waves of data from the same participants (all but Lucid, which can only recontact a subset of its participant base). Finally, the cost per participant varied substantially across samples, although these costs are quite dynamic and may have changed by the time this paper is published. Be that as it may, at the time we conducted these surveys, Bovitz-Forthright and Prolific NR were the most expensive, followed by Prolific, Connect, Connect NR, and CR Toolkit, followed by Open Mturk and CRSTAL, with Lucid as the least expensive.

Conclusion

These results provide clear evidence that numerous common sources for recruiting opt-in online research participants vary widely on response quality, representativeness, and professionalism. Thus, it is imperative that researchers - and reviewers - think carefully about the research question being asked and the quantity-of-interest being measured, as different samples are better suited for different types of research questions. There are crucial tradeoffs that need to be taken into account. That said, based on our analysis, we generally suggest the following rules of thumb:

- Researchers asking questions with low social/political focus but high requirements of attention for experiments should consider one of the samples that scored higher on response quality, lower on representativeness with more professional, experienced participants and ample platform features (e.g., Connect, Prolific, or CR Toolkit with front end filters).

- Researchers asking questions involving social/political content with lower requirements for respondent attention (e.g. shorter treatments/simpler measures) should consider one of the more representative samples containing respondents with less experience and exposure to studies (e.g. Bovitz or Lucid, or to a lesser degree Connect NR), with the addition of front end filters.
    - When representativeness is crucial, samples should not be constrained to participants using laptops or computers, as there is a strong association between mobile phone use and platform representativeness.
    - Sample weights can help less representative samples recover main effects, but are not necessarily sufficient to capture moderated effects.

- When collecting data, and critically when using the more representative samples (that often include mobile phone users), researchers should always apply attention filters on the front end of their studies.[25] Despite the variation in response quality, we were able to identify and remove inattentive participants from all samples using two simple attention checks embedded in the front-end demographics; thus, irrespective of sample, quality gains from filtering are always relevant. Further, most providers do not charge for respondents who fail attention checks placed toward the beginning of a survey. This does not address all attention concerns but can help with overall response quality.

[25] Due to response quality, a larger sample will likely be necessary on Lucid compared to the other platforms in order to achieve sufficiently precise estimates due to the lower quality data.

- Researchers asking questions involving social/political content and who have higher requirements of attention and participant naivete should consider the representative sample that was able to still obtain higher quality data (i.e. Bovitz-Forthright), but with front end filters. However, this option is also on the higher end of price for the samples tested here.

- Finally, there is no context observed here where researchers should be using open MTurk. This is the one sample that performed objectively worse than all others on all dimensions. It had the lowest representativeness, one of the worst response quality scores, some of the highest exposure and experience rates. Additionally and concerningly, participants on open Mturk showed the lowest level of reported honesty, making much of the other content provided further in question. These results, along with other recently published work (e.g., Peer et al., 2022), should clearly direct researchers away from this sample and to any of the other samples discussed.

In sum, across these nine samples, we observed broad and substantial variation on response quality, representativeness and participant naivete and experience. We find that all of these factors show interesting variation as well as clear associations. Our aim is to help researchers in the social sciences, as well as those who consume this research, be more informed and better able to evaluate results that come from such samples. It is critical that researchers understand that the participants they test and question are a crucial aspect of the answers they obtain and thus inform on the value of such answers. As we note above, different samples may be better suited to different questions, and thus great care should be given to which types of respondents are selected to participate in a given study.

References

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political analysis*, 20(3), 351-368.

Berinsky. A. J., Frydman, A., Margolis, M. F., Sances, M. W. & Valerio, D. C. (forthcoming in POQ). Measuring Attentiveness in Self-Administered Surveys.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. John Wiley & Sons.

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour*, *5*(8), 980-989.

Chaturvedi, R. (2016). A closer look at the gender gap in presidential voting.*Pew Research Center*. 7/28, accessed 2/16/24, https://www.pewresearch.org/short-reads/2016/07/28/a-closer-look-at-the-gender-gap-in-presidential-voting/

Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias?. *Public Opinion Quarterly*, *79*(3), 790-802.

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology?. *Research & Politics*, *2*(4), 2053168015622072.

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49), 12441-12446.

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, 7(3), 613-628.

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*, 6(1), 2053168018822174.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., ... & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4-36.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos one*, *18*(3), e0279720.

Druckman, J. N. (2001). Using credible advice to overcome framing effects. *Journal of Law, Economics, and Organization*, 17(1), 62-82.

Druckman, J. N., & Kam, C. D. (2011). Students as experimental participants. *Cambridge handbook of experimental political science*, 1, 41-57.

Druckman, J. N. (2022). *Experimental thinking*. Cambridge University Press.

Enns, P., & Rothschild, J. (2021). Revisiting the 'Gold Standard' of Polling: New Methods Outperformed Traditional Ones in 2020. *Medium* [blog], March, 18.

Gelman, Andrew. 2018. “You Need 16 Times the Sample Size to Estimate an Interaction than to Estimate a Main Effect.” Statistical Modeling, Causal Inference, and Social Science. Retrieved March 23, 2024

Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public opinion quarterly*, 76(3), 491-511.

Halevy, R., Shalvi, S., & Verschuere, B. (2014). Being honest about dishonesty: Correlating self-reports and actual lying. Human Communication Research, 40(1), 54-72.

Hillygus, D. S., Jackson, N., & Young, M. (2014). Professional respondents in nonprobability online panels. *Online Panel Research: Data Quality Perspective, A*, 219-237.

Hillygus, D. S., & LaChapelle, T. (2022). Diagnosing survey response quality. *Handbook on Politics and Public Opinion*, 10-25.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14, 399-425.

Jefferson, H. (2020). The curious case of Black conservatives: construct validity and the 7-point liberal-conservative scale. *Available at SSRN 3602209*.

Jerit, J., & Barabas, J. (2023). Are Nonprobability Surveys Fit for Purpose?. *Public Opinion Quarterly*, 87(3), 816-840.

Kahan, D. M., & Peters, E. (2017). Rumors of the 'Nonreplication' of the 'Motivated Numeracy Effect' are greatly exaggerated. *Yale Law & Economics Research Paper*, (584).

Kahan, D. M. (2015). The politically motivated reasoning paradigm. *Emerging Trends in Social & Behavioral Sciences*, Forthcoming.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk, *Econometrica* 47, 263-291.

Kam, C. D., & Simas, E. N. (2010). Risk orientations and policy frames. *The Journal of Politics*, 72(2), 381-396.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of advertising*, *46*(1), 141-155.

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., & Asare-Marfo, D. (2021). Strategies for detecting insincere respondents in online polling. *Public Opinion Quarterly*, *85*(4), 1050-1075.

Krefeld-Schwalb, A., Sugerman, E. R., & Johnson, E. J. (2024). Exposing omitted moderators: Explaining why effect sizes differ in the social sciences. *Proceedings of the National Academy of Sciences*, *121*(12), e2306281121.

Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational behavior and human decision processes*, 75(1), 23-55.

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M. J. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, *82*(4), 707-744.

Malhotra, N., & Popp, E. (2012). Bridging partisan divisions over antiterrorism policies: The role of threat perceptions. *Political Research Quarterly*, 65(1), 34-47.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., & Campos, L. F. (2018). Worth weighting? How to think about and use weights in survey experiments. *Political Analysis*, *26*(3), 275-291.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, 2(2), 109-138.

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton University Press.

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22-27.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5), 411-419.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). *Social heuristics shape intuitive cooperation*. Nature communications, 5(*1*), 3677.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Academic Press.

Smith, T. W. (1987). That which we call welfare by any other name would smell sweeter an analysis of the impact of question wording on response patterns. *Public opinion quarterly*, 75-83.

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184-197.

Tipton, E., Yeager, D. S., Iachan, R., & Schneider, B. (2019). Designing probability samples to study treatment effect heterogeneity. *Experimental methods in survey research: Techniques that combine random sampling with random assignment*, 435-456.

Valentino, N. A., Zhirkov, K., Hillygus, D. S., & Guay, B. (2020). The consequences of personality biases in online panels for measuring public opinion. *Public Opinion Quarterly*, *84*(2), 446-468.

Supplementary Information

Representativeness Versus Attentiveness: A Comparison Across Nine Online Survey Samples

M.N. Stagnaro, J.N. Druckman, A.J. Berinsky, A.A. Arechar, R. Willer and D.G. Rand

Contents

Supplementary Section	Page number
SI-1 Specifics of each Sample	p26
SI-2 Survey Materials	p26
SI-3 Probability Sample Information	p26
SI-4 Demographics Broken Down by Sample with Benchmark Comparison	p27
SI-5 Compensation on response quality	p30
SI-6 Attention Check Wording	p32
SI-7 Response Quality Breakdown for Top Five Samples	p33
SI-8 Response Quality and Demographic Representativeness	p33
SI-9 Response Quality and Typicality Representativeness	p38
SI-10 Helping the Poor policy frame - Main effect	p42
SI-11 Loss Aversion Study Details	p42
SI-12 Loss Aversion Main effects by Sample	p43
SI-13 Loss Aversion Interaction effects by Sample	p43
SI-14 Exhaustive Comparison Analysis	p45
SI-15 Mobile Phone Use and Response Quality	p48
SI-16 Lying and Honesty Across Samples	p49
SI-17 Effect of filtering on treatment effects	p52
SI-18 The Relationship Between Overall Representativeness and Response Quality	p56
SI-19 Relationship between Duration, Representativeness, Response Quality and Attention	p57
SI- 20 Sample Profile	p59
SI-21 Post stratification weighting and treatment effects	p62
SI-22 Supplementary Information References	p64

SI-1 Specifics of each Sample

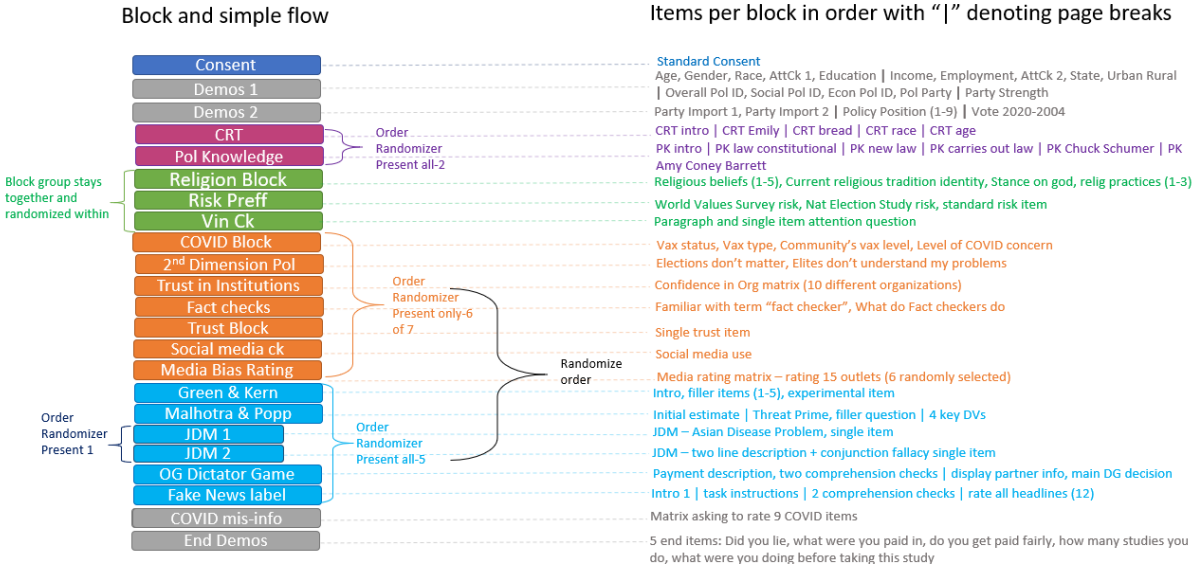
We examined nine different samples from Mturk, Prolific, CloudResearch's Connect panel, Lucid Marketplace (owned by Cint), and Bovitz-Forthright. All respondents from each sample completed the same survey. The table below provides a summary of the dates when data were collected, the sample sizes, median time taken to finish the last question of the survey (regardless of attention check passage), and the platform's public url.

Sample	Dates	N	Median time to complete	Platform's URL
Bovitz Forthright	6/8/22 – 6/20/22	1,115	24:02	https://www.forthrightaccess.com/
Lucid	6/8/22 – 7/1/22	4,505	20:28	https://luc.id/log-in-3/
Mturk				
Open Mechanical Turk	7/22/22 – 9/15/22	1,103	15:18	https://www.mturk.com/
CloudResearch's Toolkit	9/14/22 – 9/15/22	1,088	16:32	https://www.cloudresearch.com/products/turkprime-mturk-toolkit/
Stanford CRSTAL MTurk panel	6/8/22 – 6/16/22	1,077	18:11	NA ('in house" platform)
Connect				
CloudResearch Connect	9/6/22 – 9/17/22	1,046	17:28	https://www.cloudresearch.com/products/connect-for-researchers/
Connect NR	10/18/23 - 10/20/23	1,028	19:46	
Prolific				
Prolific	6/8/22 – 6/16/22	1,051	18:29	https://www.prolific.com/
Prolific NR	1/12/23 – 1/24/23	1,040	17:51	
Total	**6/8/22 – 10/20/23**	**13,053**	**18:50**	

SI Table 1-1. *Specifics of each of the nine different samples studied.*

SI- 2 Survey Materials

The Figure below shows the ordering and flow of survey content, with items on the right.



SI Figure 2-1. *Survey blocks and items.*[26]

SI- 3 Probability Sample Information (for benchmarks)

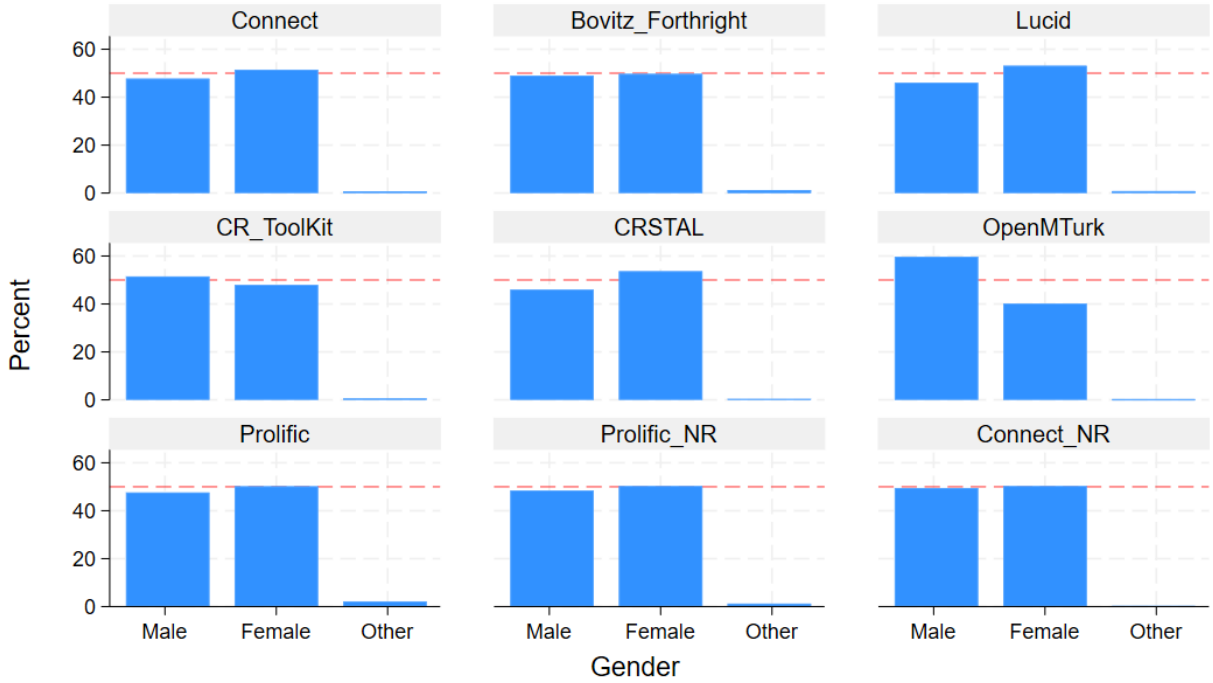
Category	Year	Source	Source's URL
Political Party affiliation	2020	ANES	https://electionstudies.org/data-tools/anes-guide/anes-guide.html?chart=party_identification_7_pt
Age	2020	ANES	https://electionstudies.org/data-tools/anes-guide/anes-guide.html?chart=age_cohort
Education	2020	ANES	https://electionstudies.org/data-tools/anes-guide/anes-guide.html?chart=education
Race	2022	Census	https://www.census.gov/quickfacts/fact/table/US/
Income	2021	NORC	NA (personally collected data)
Religious tradition	2020	ANES	https://electionstudies.org/data-tools/anes-guide/anes-guide.html?chart=religion_6_cat
Political Ideology	2020	ANES	https://electionstudies.org/data-tools/anes-guide/anes-guide.html?chart=lib_con_identification_7_pt
Military Spending	2020	ANES	https://electionstudies.org/data-tools/anes-guide/anes-guide.html?chart=increase_decrease_military_spending_7_pt
Gun Control	2020	ANES	https://electionstudies.org/data-tools/anes-guide/anes-guide.html?chart=gun_purchase_control
Death penalty	2020	ANES	https://electionstudies.org/data-tools/anes-guide/anes-guide.html?chart=death_penalty
Immigration	2020	ANES	https://electionstudies.org/data-tools/anes-guide/anes-guide.html?chart=immigrants_should_be_decreased_increased
Trust in Federal Gov.	2020	ANES	https://electionstudies.org/data-tools/anes-guide/anes-guide.html?chart=trust_fed_govt_5_pt
Belief in God	2014	PEW	https://www.pewresearch.org/religion/religious-landscape-study/compare/belief-in-god/by/party-affiliation/
Voted 2016	2021	NORC	NA (personally collected data)
Trust in People	2022	GSS	https://gss.norc.org/get-the-data/stata

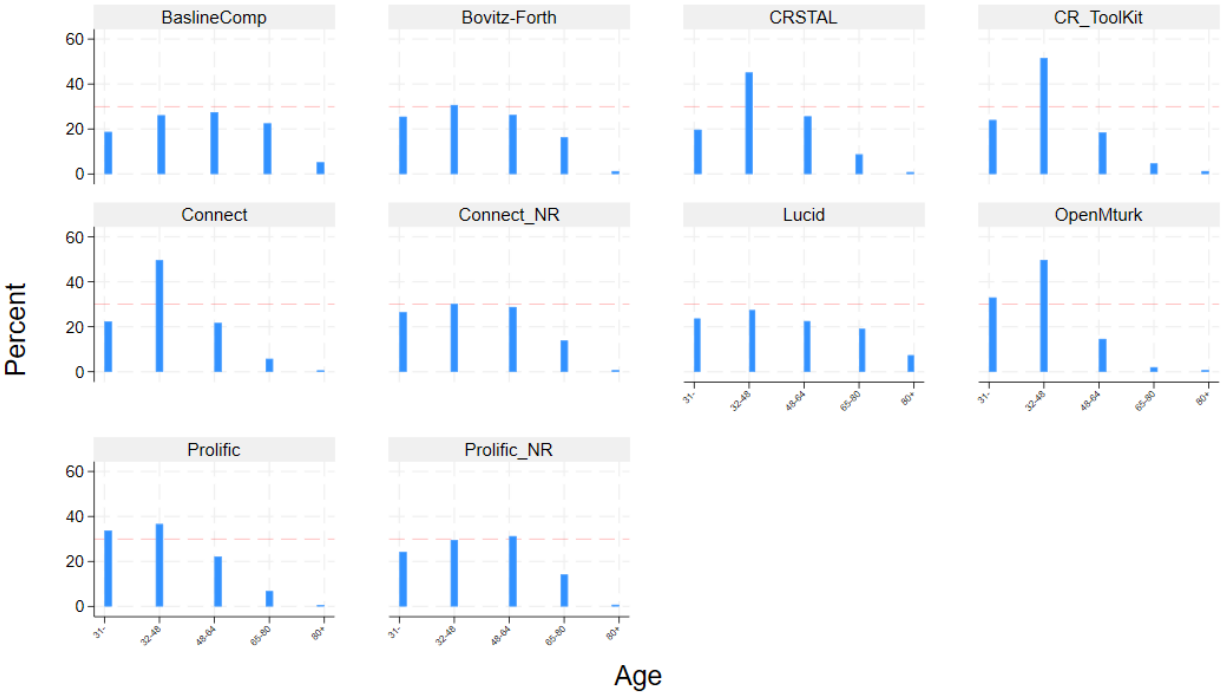
[26] Note: Not all content included in the survey was for this project. We include all content for transparency.

SI Table 3-2. *Probability sample information. Notes: the race data at the given link has been updated to the most recent data (e.g., 2022) since we used it (the 2020 data); military spending and immigration exclude "I don't know" from options.*

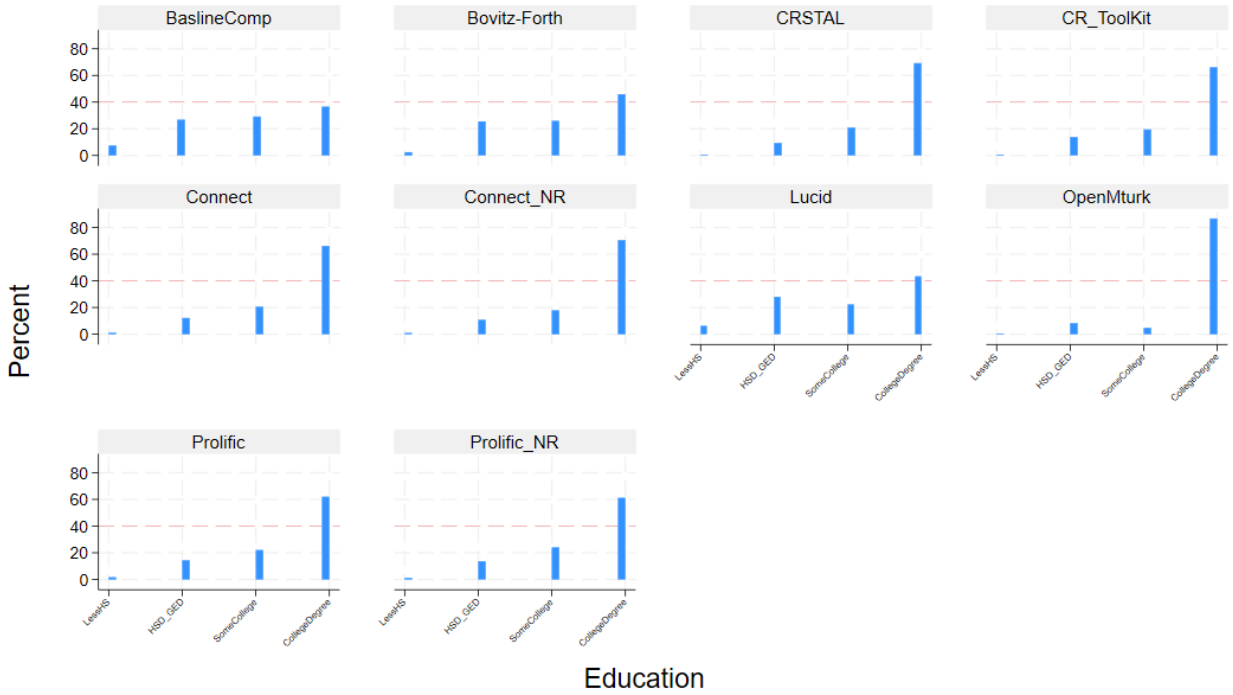
SI- 4 Demographics Broken Down by Sample with Benchmark Comparison

SI Figure 4-0. *Distribution of gender for all samples. Red dashed line indicates the 50% mark. Note that gender was not included as one of the representative dimensions due to only minor variation.*



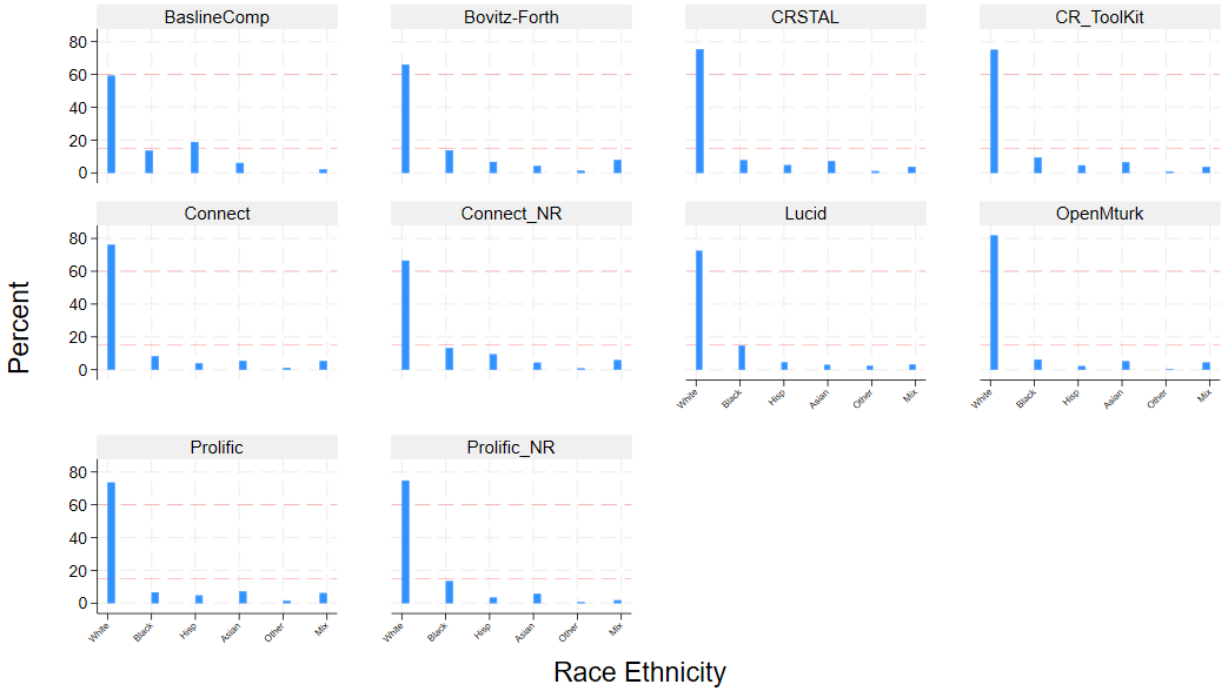


SI Figure 4-1. *Distribution of age for all samples compared to the probability ("Benchmark") sample, top left. Red dashed line for aiding readers with making cross sample comparisons.*

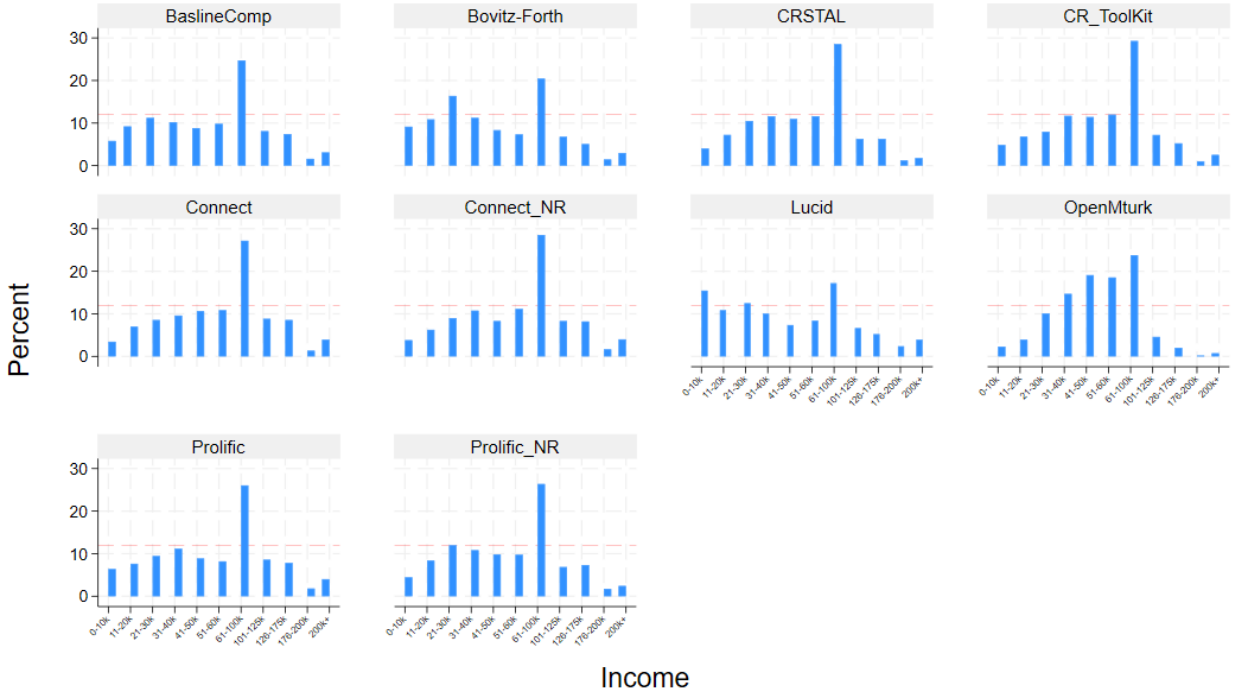


SI Figure 4-2. *Distribution of education for all samples compared to the probability ("Benchmark") sample, top left. Red dashed line for aiding readers with making cross sample comparisons.*

SI Figure 4-3. *Distribution of political party affiliation, and strength, for all samples compared to the probability ("Benchmark") sample, top left. Red dashed line for aiding readers with making cross sample comparisons.*



SI Figure 4-4. *Distribution of race/ ethnicity for all samples compared to the probability ("Benchmark") sample, top left. Red dashed line for aiding readers with making cross sample comparisons.*



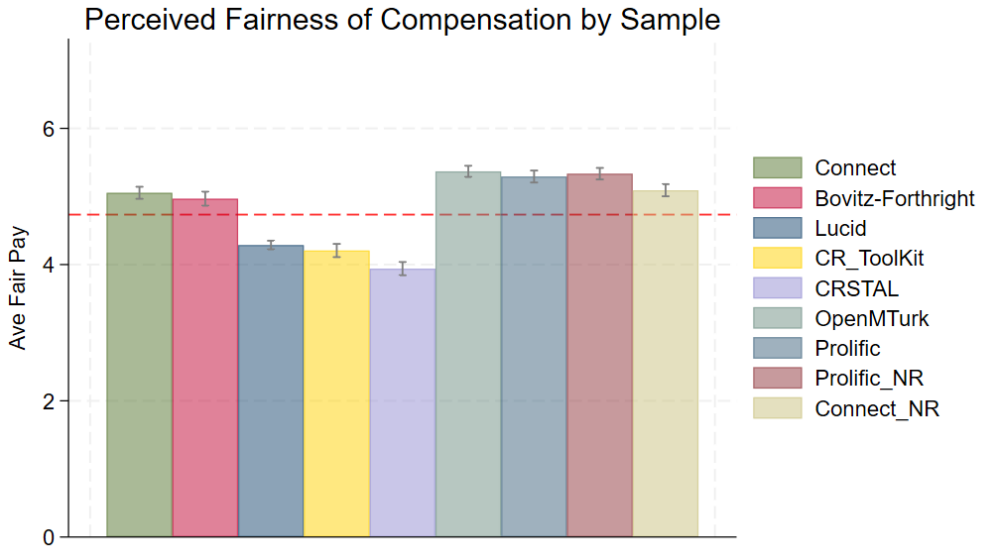
SI Figure 4-5. *Distribution of income for all samples compared to the probability ("Benchmark") sample, top left. Red dashed line for aiding readers with making cross sample comparisons.*

SI Figure 4-6. *Distribution of religious tradition for all samples compared to the probability ("Benchmark") sample, top left. Red dashed line for aiding readers with making cross sample comparisons.*

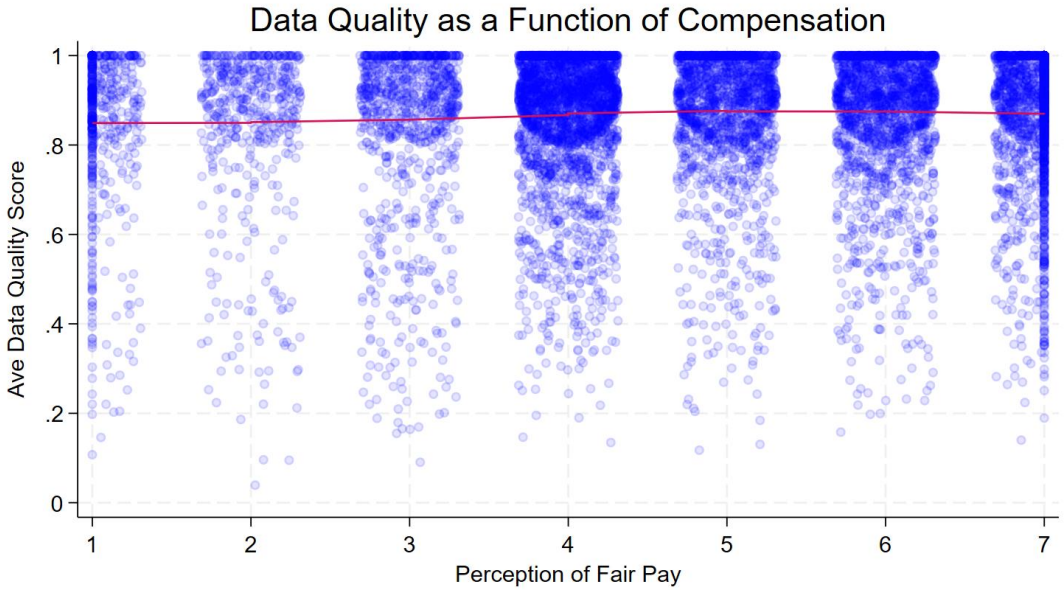
SI- 5 Effect of compensation on response quality

We look at perceptions of platform compensation and its potential role in response quality. We asked participants to rate if they feel they regularly get fair compensation on this platform, using a seven-point scale (0=not at all, 3=somewhat, 6=very much). Those participants obtained via Mturk are likely rating Mturk's platform rather than those panels that connect them with requesters, i.e., the CRSTAL panel, CloudResearch Toolkit, Lucid (which aggregates participants from many sources). Still, this provides some insight into the personal experience of those who are taking these studies and how they feel toward the compensation for the work they do.

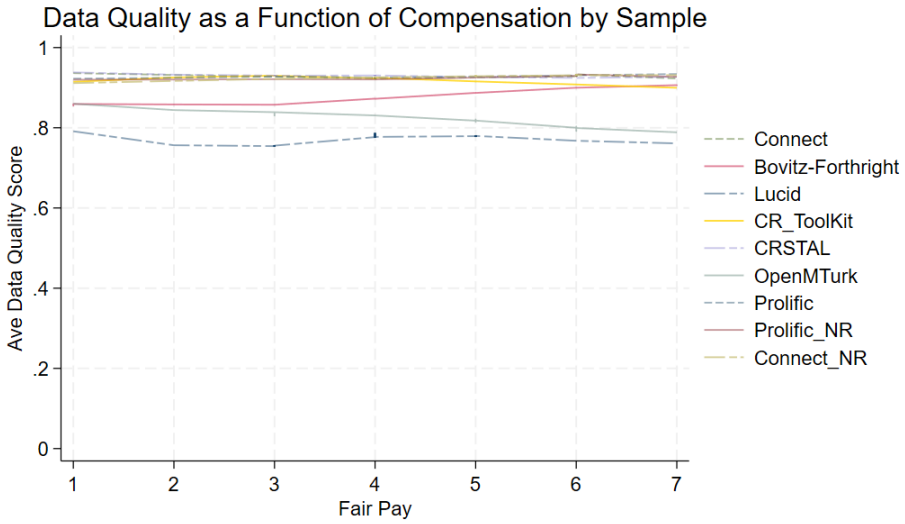
Across panels we see only small variation in perceptions of compensation (SI Figure 3a), with Open Mturk (M=5.37, SE=.04), both Prolific platforms (Prolific: M=5.29, SE=.05, Prolific_NR: M=5.33, SE=.04), and both Connect platforms (Connect: M=5.05, SE=.05; Connect_CR: M=5.09, SE=.05) on average rated quite high; with Bovitz-Forthright shortly behind (M=4.97, SE=.05); and Lucid (M=4.29, SE=.03) CR_Toolkit (M=4.21, SE=.05) and the CRSTAL panel (M=3.94, SE=.05) all lower, scoring at the midpoint of the scale. Somewhat surprisingly, when looking at the effect of perceived compensation on response quality, we see no evidence of a relationship (*B*=-.004, *p*=.623, *95%CI*[-.002-.001], SI Figure 3b). Though, this finding should not be taken as an endorsement for paying online workers a lower wage. We feel workers should be paid at a reasonable rate, due to legal obligations of minimum wage laws and ethical ideals.



SI Figure 5-1. *Mean of self-reported fair compensation for the work participants normally experience on this platform. Red line indicates the overall average.*

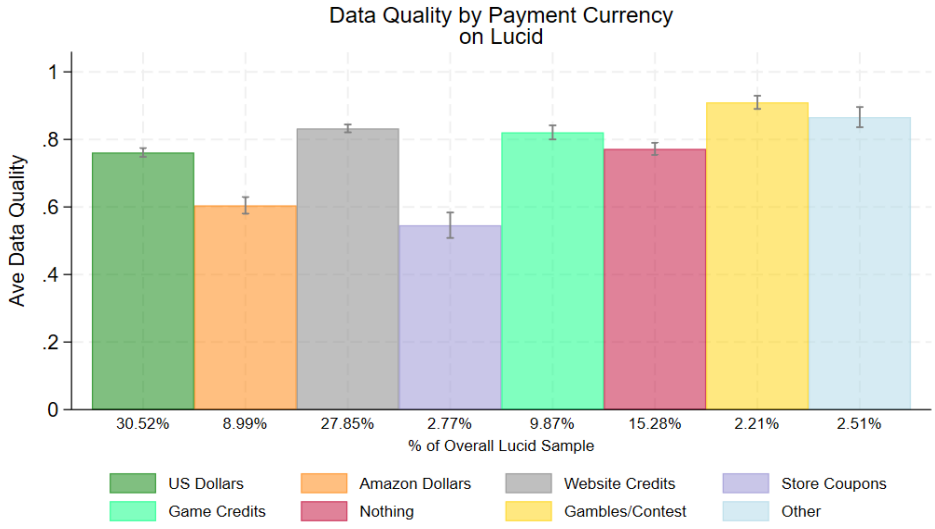


SI Figure 5-2. *Response quality as a function of perceived pay fairness, collapsing across all samples. Red line indicates a locally weighted regression line at the mean for each pay level.*



SI Figure 5-3. *Response quality as a function of perceived fairness in platform compensation for each sample. All lines indicate locally weighted regression lines for each platform in turn.*

Almost all platforms compensate respondents monetarily. However, due to its approach of aggregating from a wide set of sources, Lucid participants report being compensated in a wider array of currencies, including reporting no compensation at all. When asking participants how they were compensated for completing this study, answers included: dollars, coupons for various stores, game credits, website tokens, and no compensation at all. Interestingly, the variation in response quality across compensation types, though notable, is less than one might imagine (Figure 2). Most striking is the comparatively high response quality from participants completing the survey without any compensation. This is likely due to self-selection, where the subset of participants who are willing to participate for no compensation are internally motivated to take the studies out of interest or some other factor.



SI Figure 5-4. *Average response quality on Lucid by compensation currencies.*

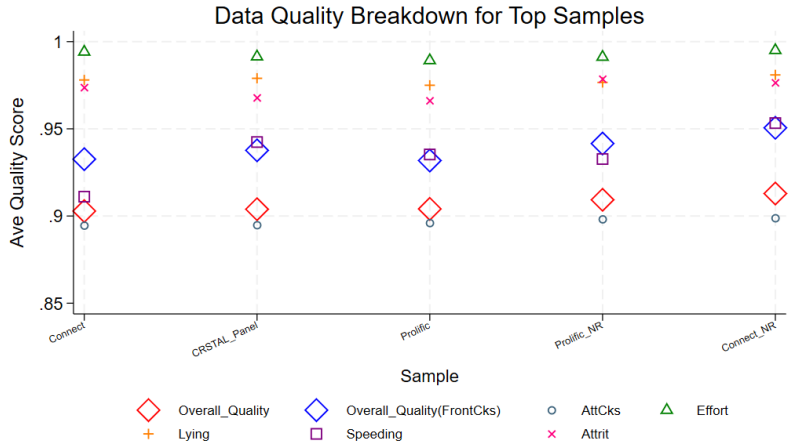
SI- 6 Attention Check Wording

There were seven attention checks distributed across the full study. All attention check items were short, easy to answer, and had no trick/deceptive aspects. These items are as follows.

1. *Please write "twenty-five" using numbers. [free response box]*
2. *Help us keep track of who is paying attention. Please select "Somewhat disagree" from the options below. [Multiple choice: Strongly agree - Strongly disagree]*
3. *Please select "Neither Oppose nor Support" [matrix options]*
4. *Please select "A great deal of confidence" [matrix options]*
5. *What type of social media accounts do you use (if any)? Please answer honestly. [do they select the option "TisFask" which does not exist nor sound like any social media option]*
6. *Please select "Accurate" in the responses provided. [matrix options]*
7. *Please ignore this question and do not answer. [choose all that apply: Strongly agree - Strongly disagree]*

SI- 7 Response Quality Breakdown for Top Five Quality Samples

Here we provide a figure looking at response quality, removing the lower scoring samples to better focus on any differences between the top five response quality samples (Connect, CRSTAL, Prolific, Prolific_NR and Connect_NR), which were highly overlapping when focusing on the full range of response quality. We focus on the subcomponents as the violin plots for this group do not meaningfully differ. As can be seen below, there is very little variation, with speeding being the most notable difference.



SI Figure 7-1. *Top five samples on response quality, focusing on the subcomponents of quality and truncating the y axes to allow better visual access to any present difference.*

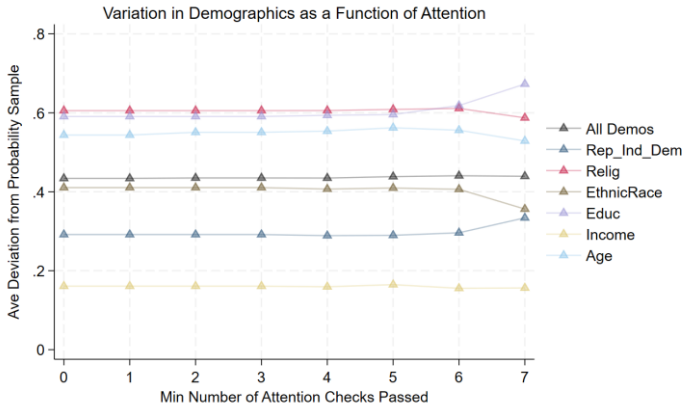
SI- 8 Response Quality and Demographic Representativeness

In the main text, we examined how increasing response quality by filtering out inattentive participants affects representativeness. We noted how there is little change in *average* demographic representativeness across all but the most extreme levels of attention filtering. However, this pattern differed somewhat when looking at the individual demographic variables, noting that the lack of change in average demographic representativeness hides a decrease in representativeness for some characteristics (e.g. education, income) and increases in representativeness for others (e.g. religiosity, age).

Below, we unpack these differences for each representativeness category by each sample, reporting how representativeness changes as a function of the minimum number of attention checks correctly answered. We first focus on demographics, then on the attitudes and beliefs for typicality. Most notably, the reader will notice that Lucid shows notable swings in representation and can only hold a high representativeness score if the filtering is not too extreme, where the other platforms can manage slightly higher levels of filtering. Below we focus on each demographic category with black representing the average for all.

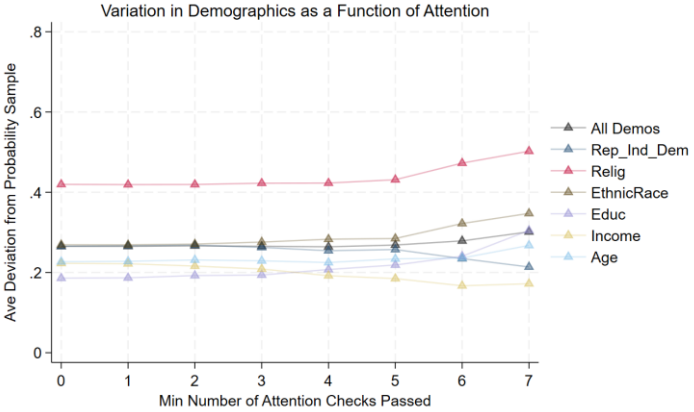
Si figure 8-10 showcases representativeness across several other dimensions. Specifically we look at, effort spent (1= effort was taken to answer correctly, 0= not), level of reported honesty (3= they were honest, 2= they could not recall, 1= lied at some point), and lastly, speeding (0= finished in the bottom 10% of times, 1= above the bottom 10% of completion times).

Connect:



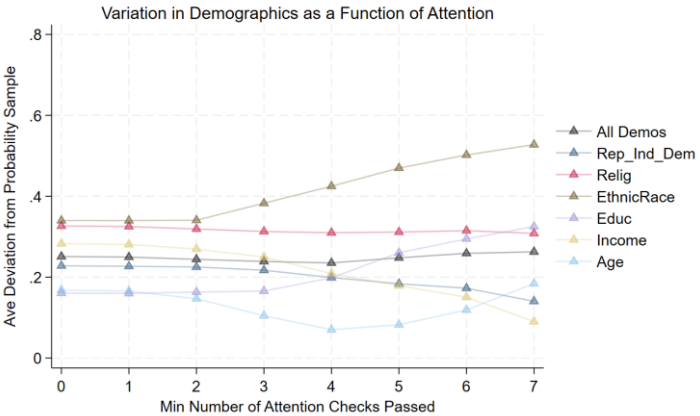
SI figure 8-1. *Representativeness for each demographic broken down by level of attention check passed, for the sample Connect.*

Bovitz:



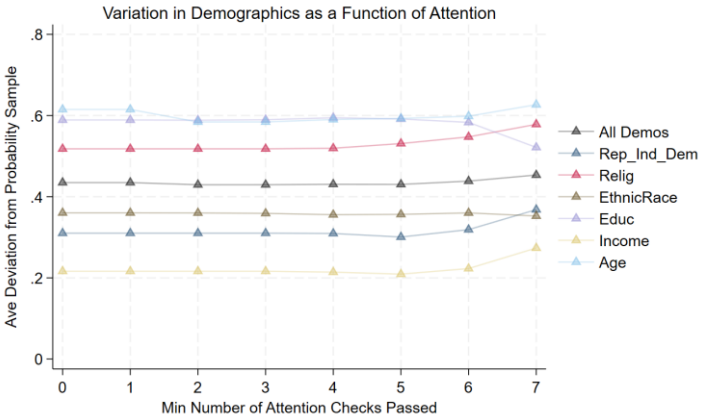
SI figure 8-2. *Representativeness for each demographic broken down by level of attention check passed, for the sample Bovitz.*

Lucid:



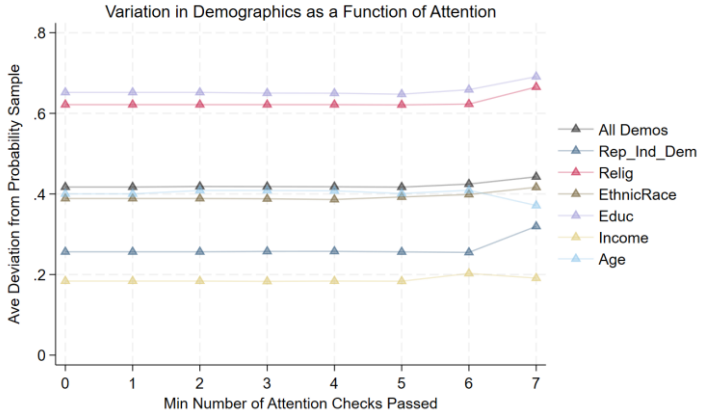
SI figure 8-3. *Representativeness for each demographic broken down by level of attention check passed, for the sample Lucid.*

CR_ToolKit:



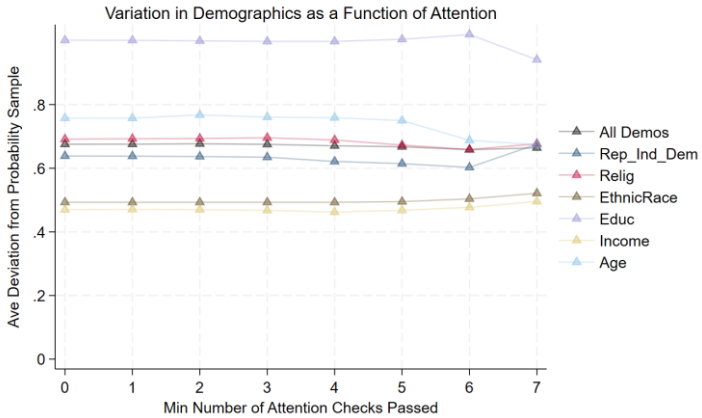
SI figure 8-4. *Representativeness for each demographic broken down by level of attention check passed, for the sample CR Toolkit.*

CRSTAL:



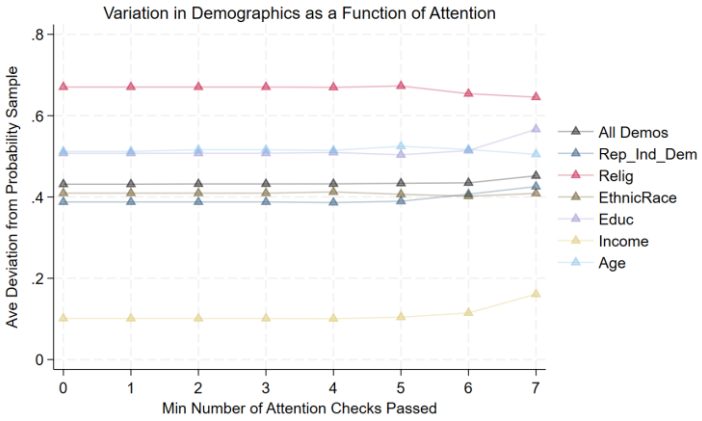
SI figure 8-5. *Representativeness for each demographic broken down by level of attention check passed, for the sample CRSTAL.*

Open Mturk:



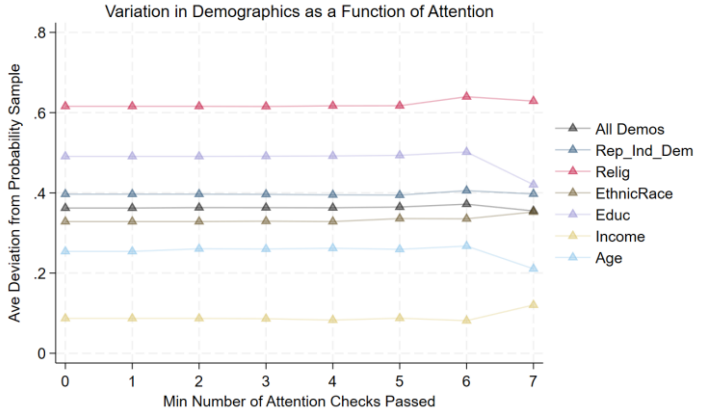
SI figure 8-6. *Representativeness for each demographic broken down by level of attention check passed, for the sample Open Mturk.*

Prolific:



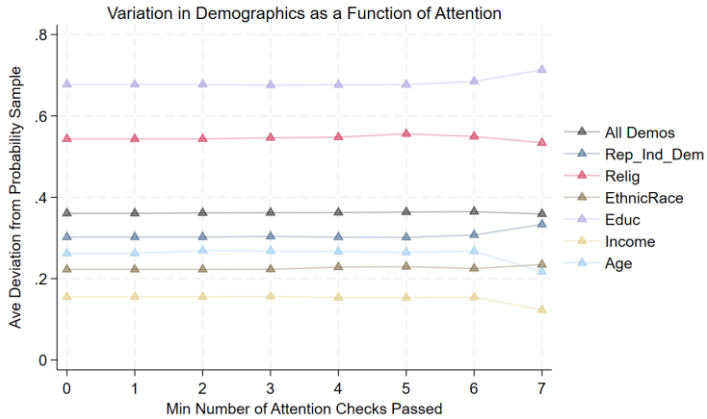
SI figure 8-7. *Representativeness for each demographic broken down by level of attention check passed, for the sample Prolific.*

Prolific_NR:



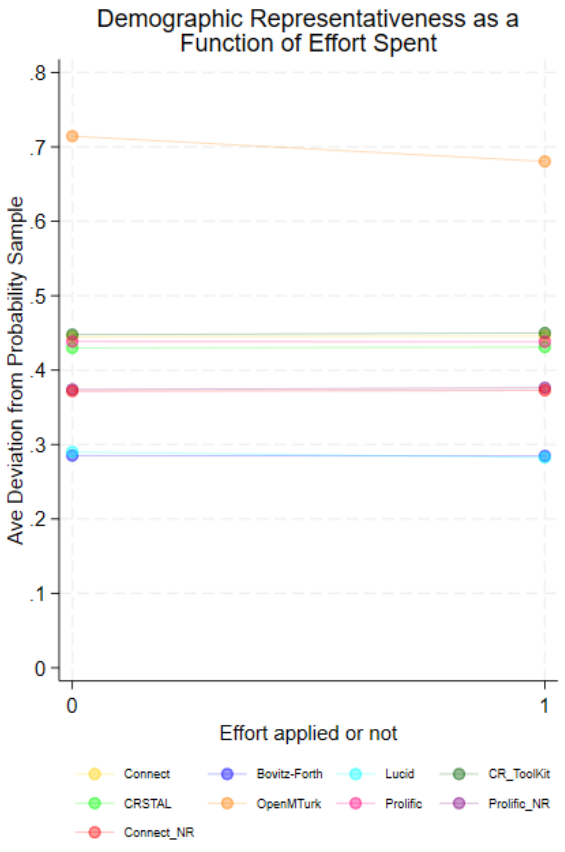
SI figure 8-8. *Representativeness for each demographic broken down by level of attention check passed, for the sample Prolific NR.*

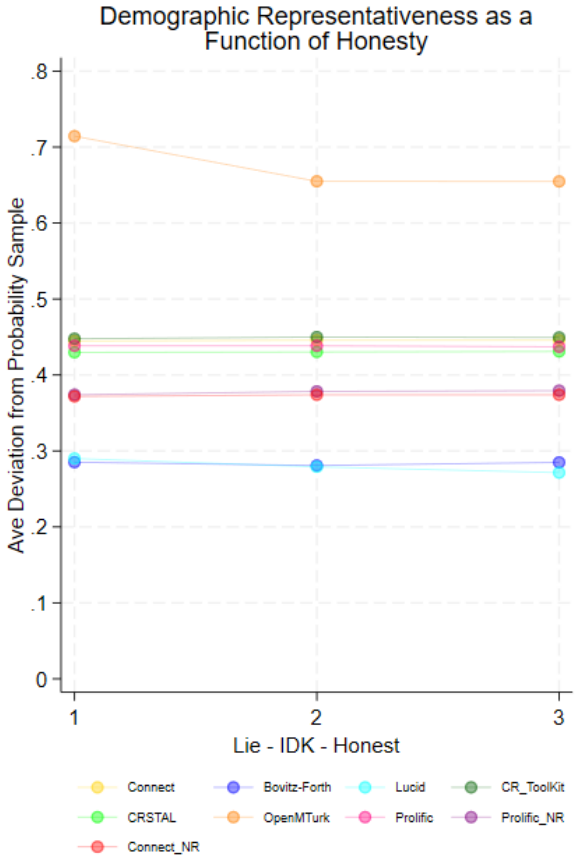
Connect_NR:

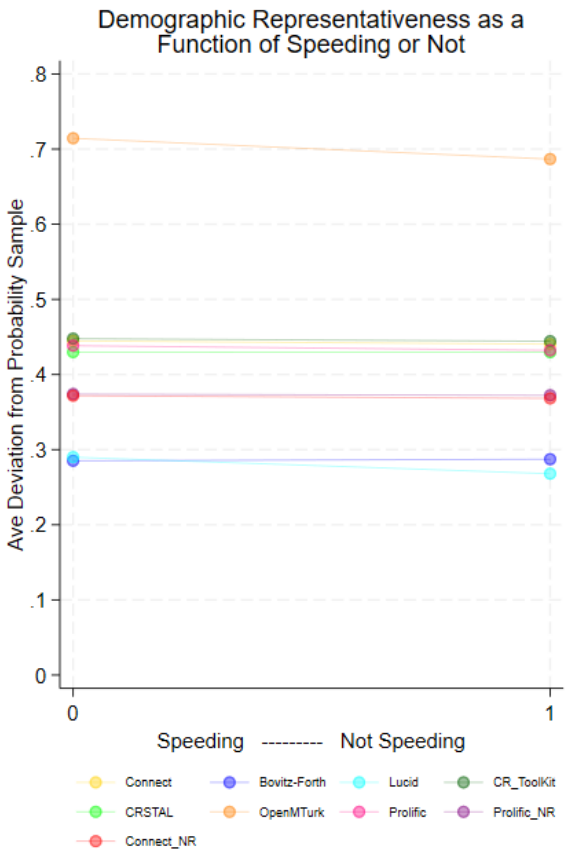


SI figure 8-9. *Representativeness for each demographic broken down by level of attention check passed, for the sample Connect NR.*

SI Figure 8-10. *Relationship between representativeness and the other response quality items*


Demographic Representativeness as a Function of Effort Spent
Ave Deviation from Probability Sample
.8
.7
.6
.5
.4
.3
.2
.1
0
0
1
Effort applied or not
Connect
Bovitz-Forth
Lucid
CR_ToolKit
CRSTAL
OpenMTurk
Prolific
Prolific_NR
Connect_NR

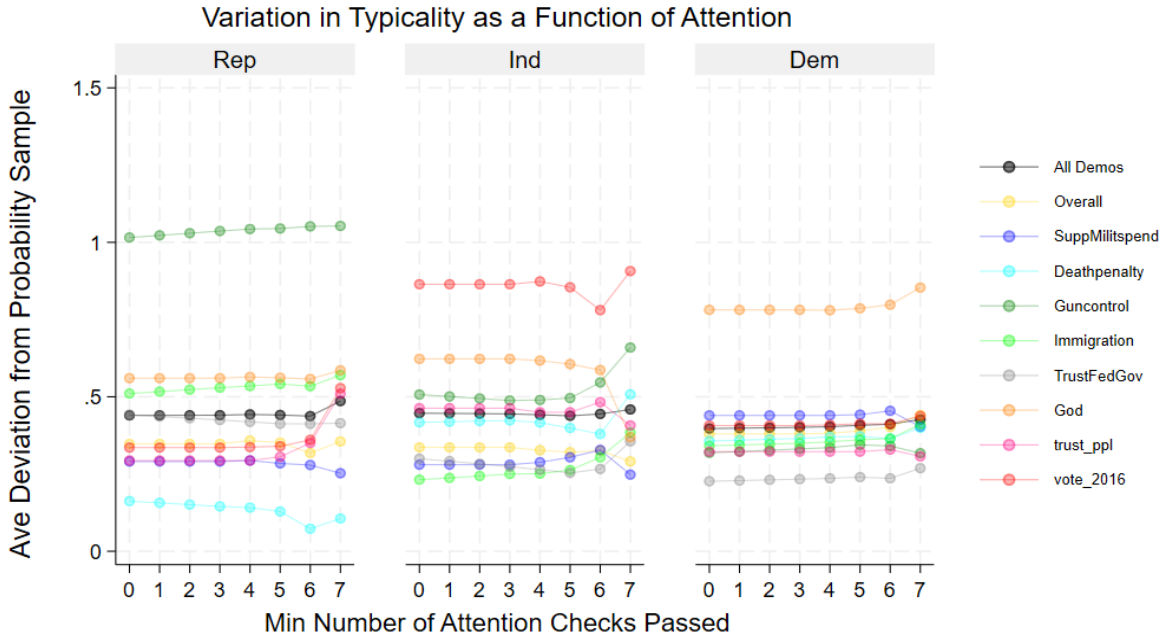

Demographic Representativeness as a Function of Honesty
Ave Deviation from Probability Sample
.8
.7
.6
.5
.4
.3
.2
.1
0
1
2
3
Lie - IDK - Honest
Connect
Bovitz-Forth
Lucid
CR_ToolKit
CRSTAL
OpenMTurk
Prolific
Prolific_NR
Connect_NR


Demographic Representativeness as a Function of Speeding or Not
Ave Deviation from Probability Sample
.8
.7
.6
.5
.4
.3
.2
.1
0
0
1
Speeding --------- Not Speeding
Connect
Bovitz-Forth
Lucid
CR_ToolKit
CRSTAL
OpenMTurk
Prolific
Prolific_NR
Connect_NR

SI-9 Response Quality and Typicality Representativeness

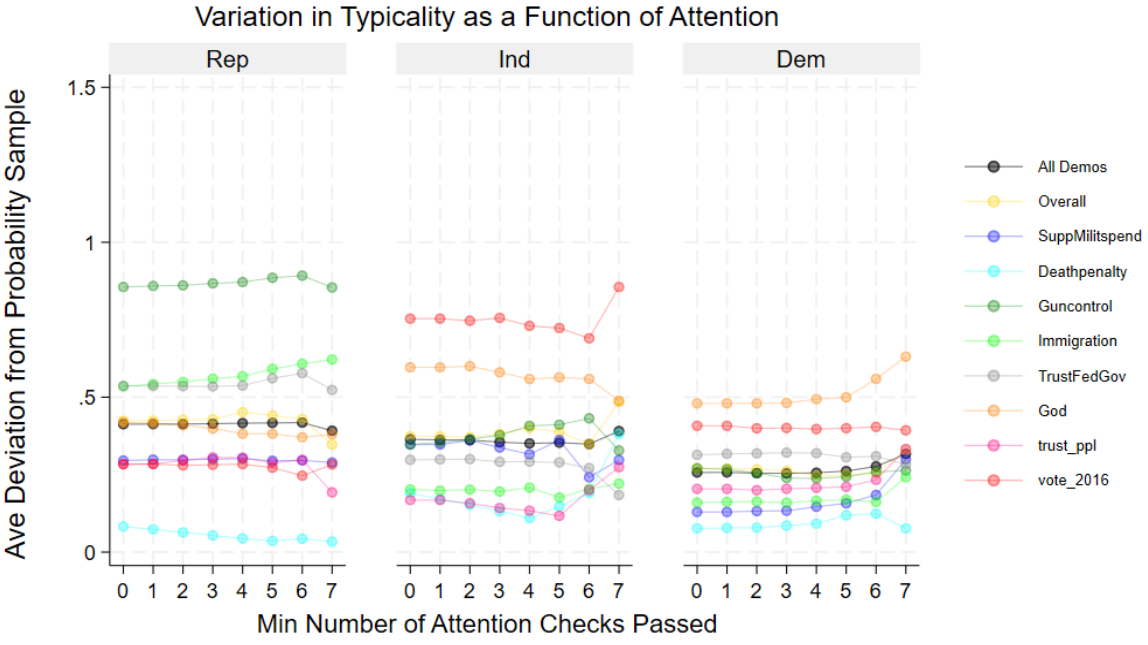
As we showcased changes in demographic representativeness above, here we focus on typicality representativeness, with black again representing the average. As in the main paper, results are broken out by political party affiliation and the minimum number of attention checks correctly answered.

Connect:



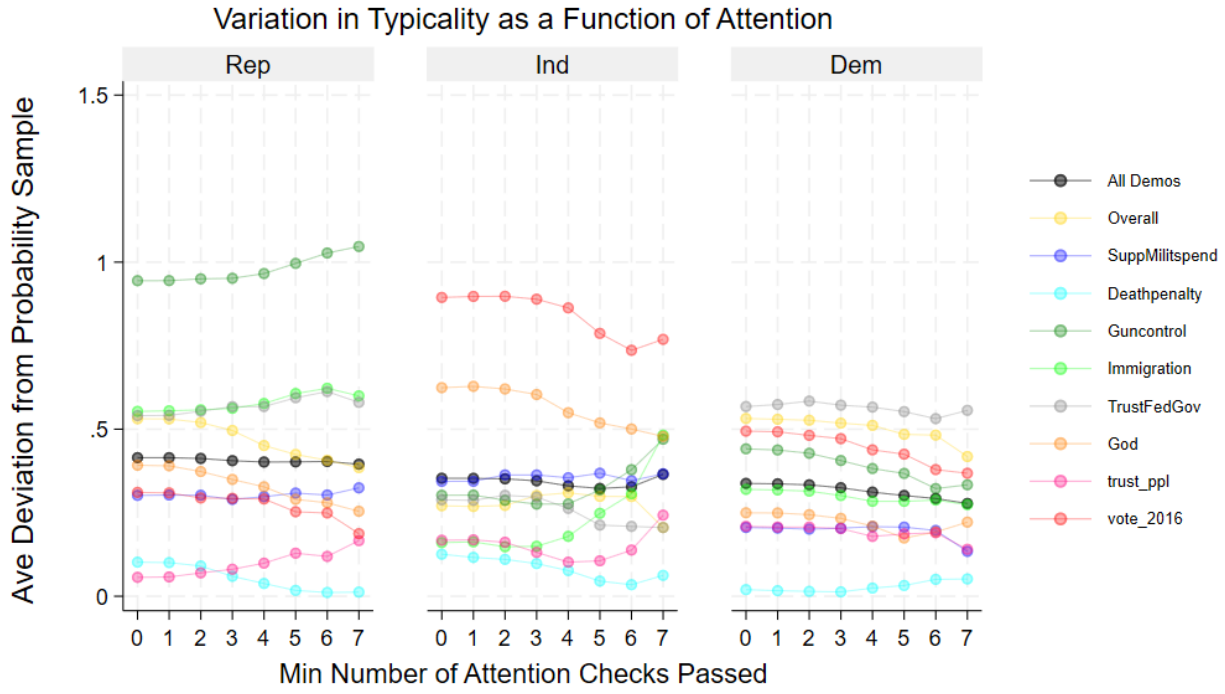
SI figure 9-1. *Representativeness for each Characteristic broken down by political party affiliation and level of attention check passed, for the sample Connect.*

Bovitz:



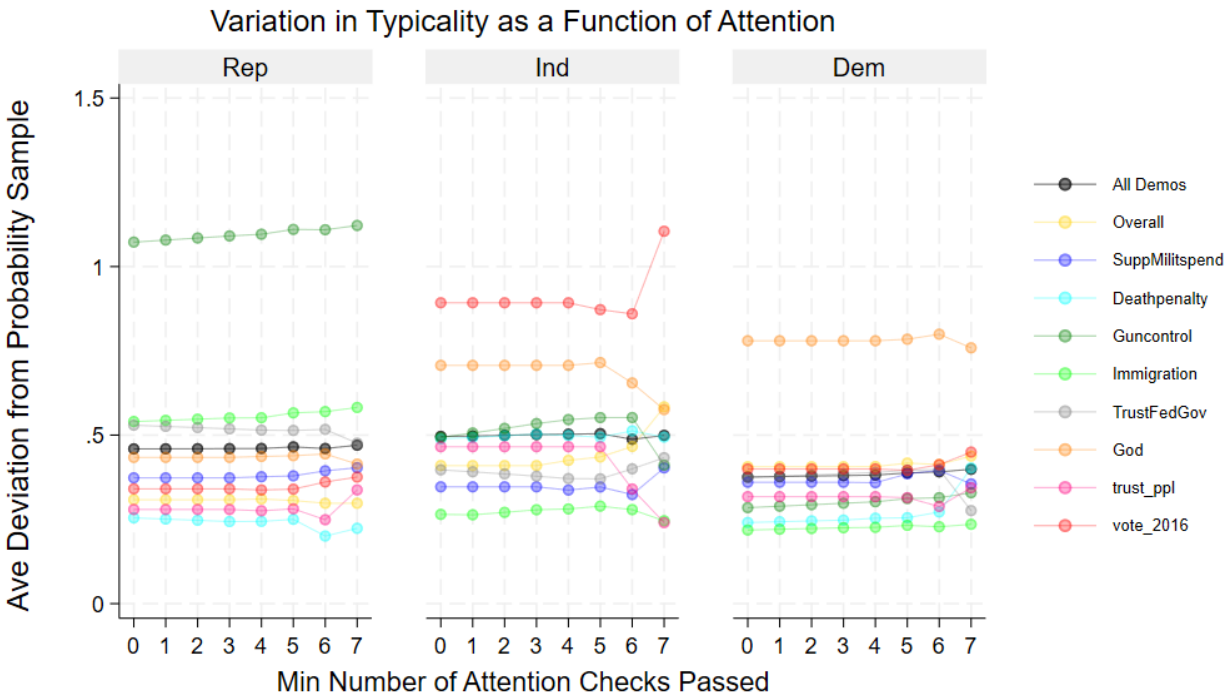
SI figure 9-2. *Representativeness for each Characteristic broken down by political party affiliation and level of attention check passed, for the sample Bovitz.*

Lucid:



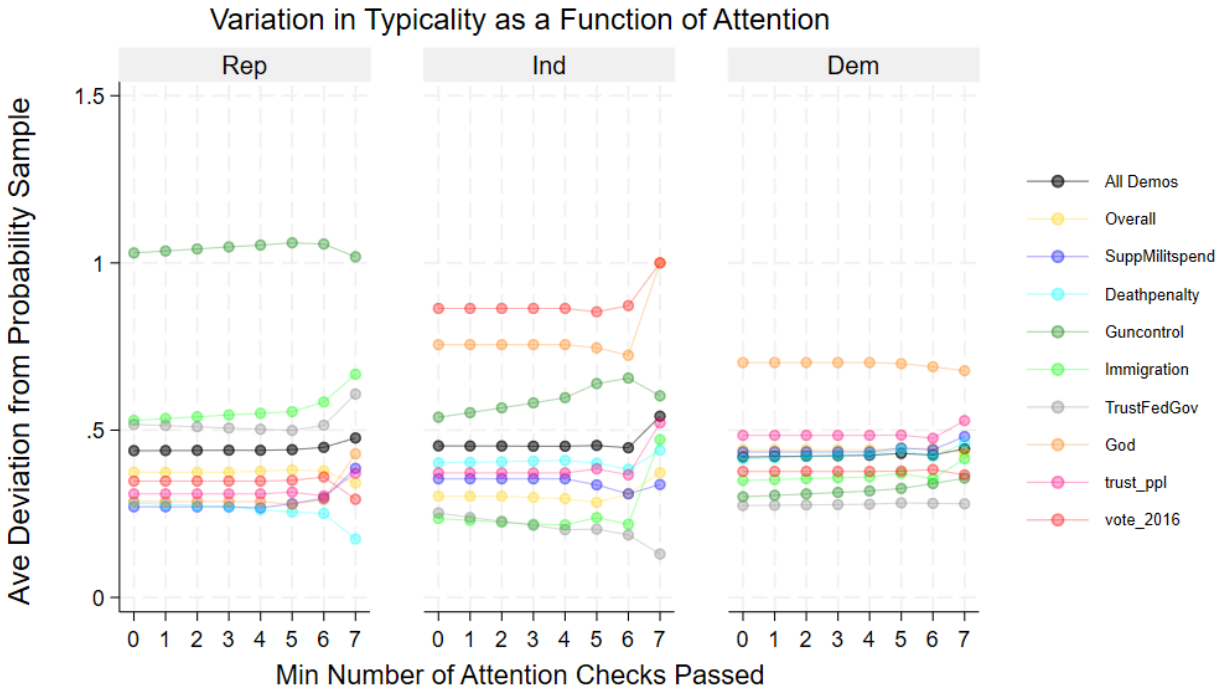
SI figure 9-3. *Representativeness for each Characteristic broken down by political party affiliation and level of attention check passed, for the sample Lucid.*

CR_ToolKit:



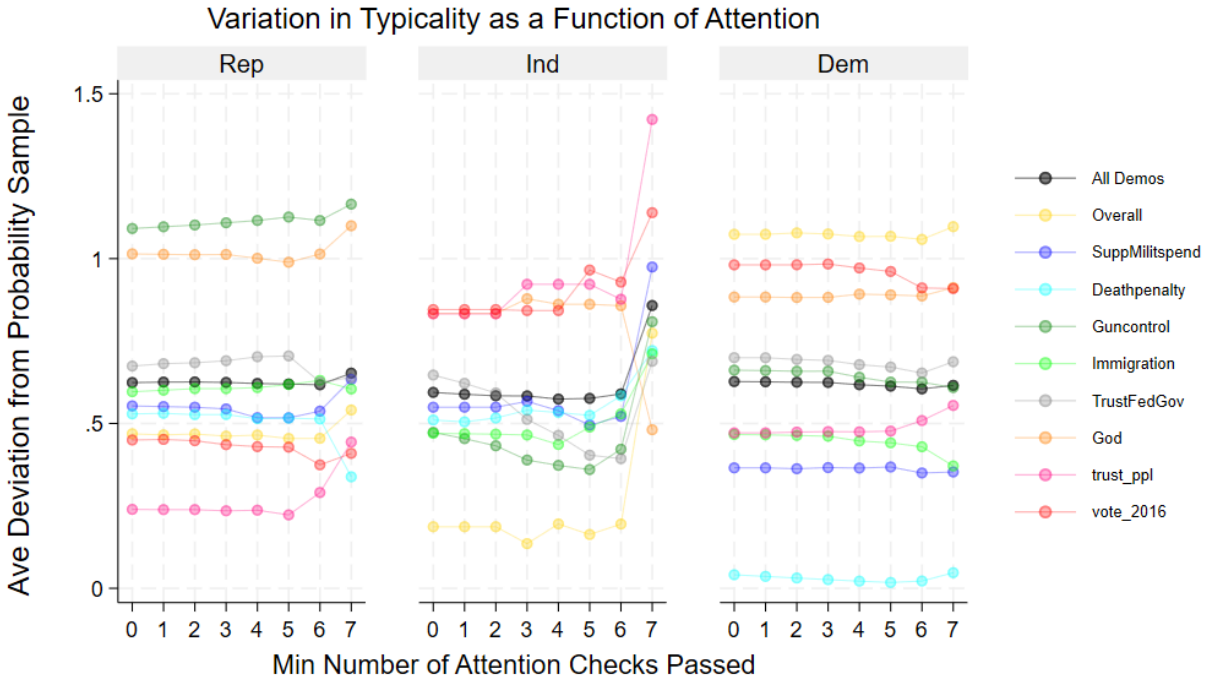
SI figure 9-4. *Representativeness for each Characteristic broken down by political party affiliation and level of attention check passed, for the sample CR Toolkit.*

CRSTAL:



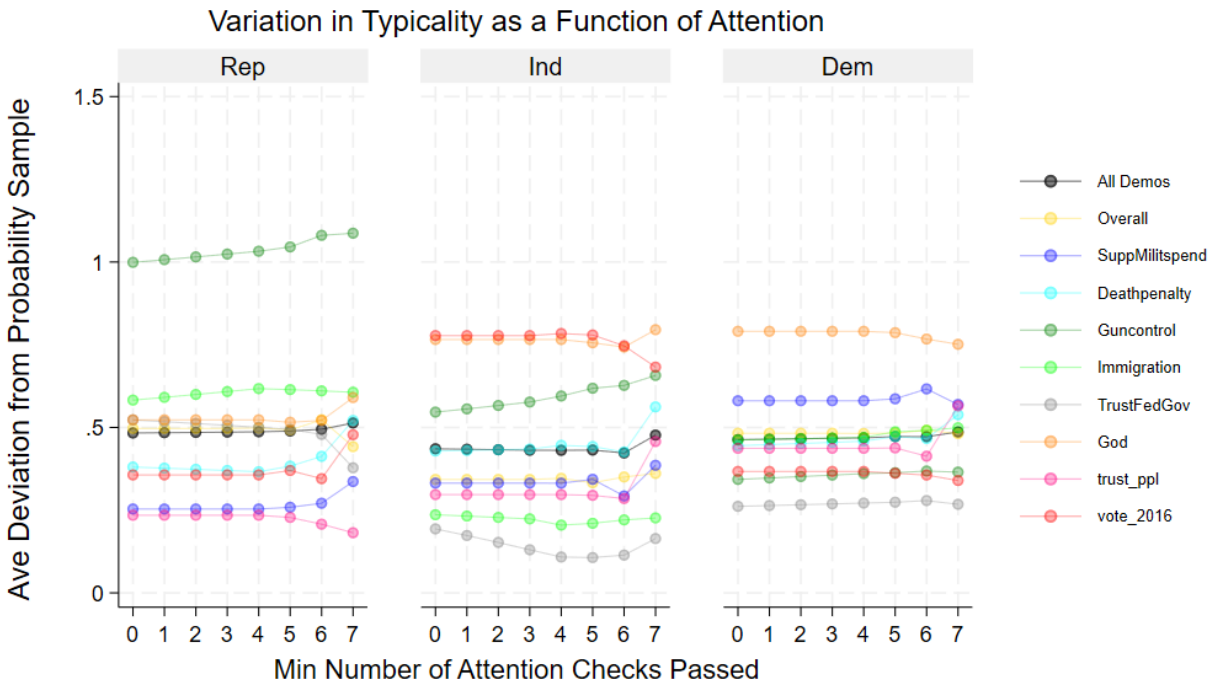
SI figure 9-5. *Representativeness for each Characteristic broken down by political party affiliation and level of attention check passed, for the sample CRSTAL.*

Open Mturk:



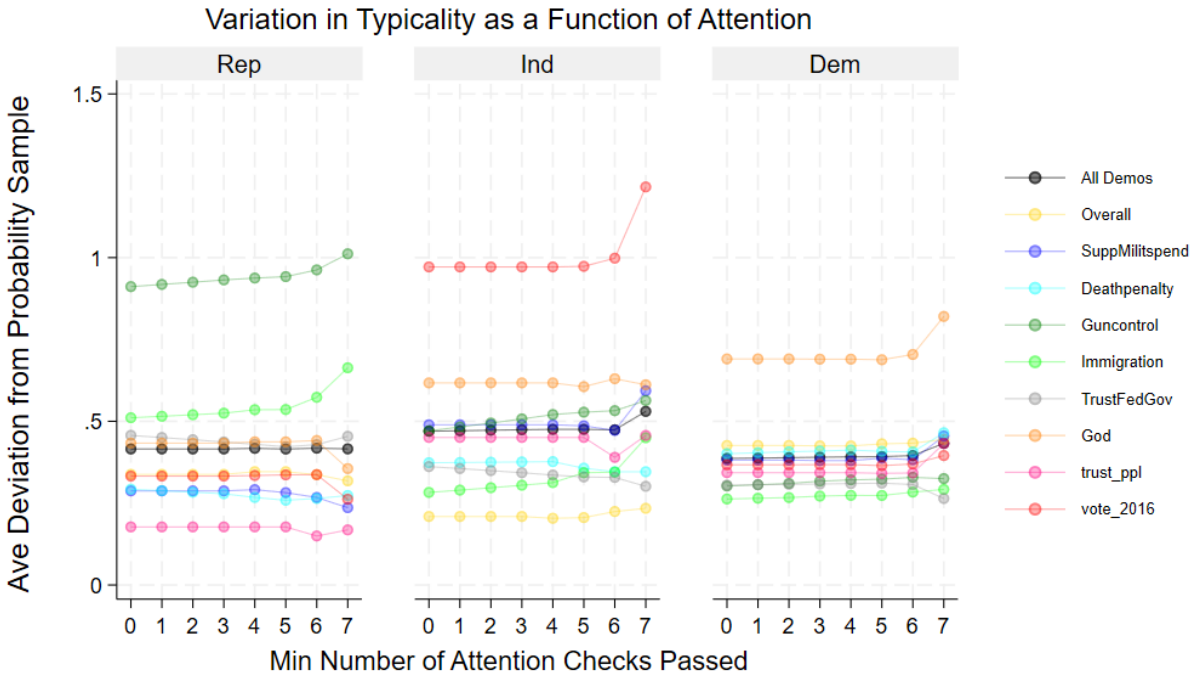
SI figure 9-6. *Representativeness for each Characteristic broken down by political party affiliation and level of attention check passed, for the sample Open Mturk.*

Prolific:



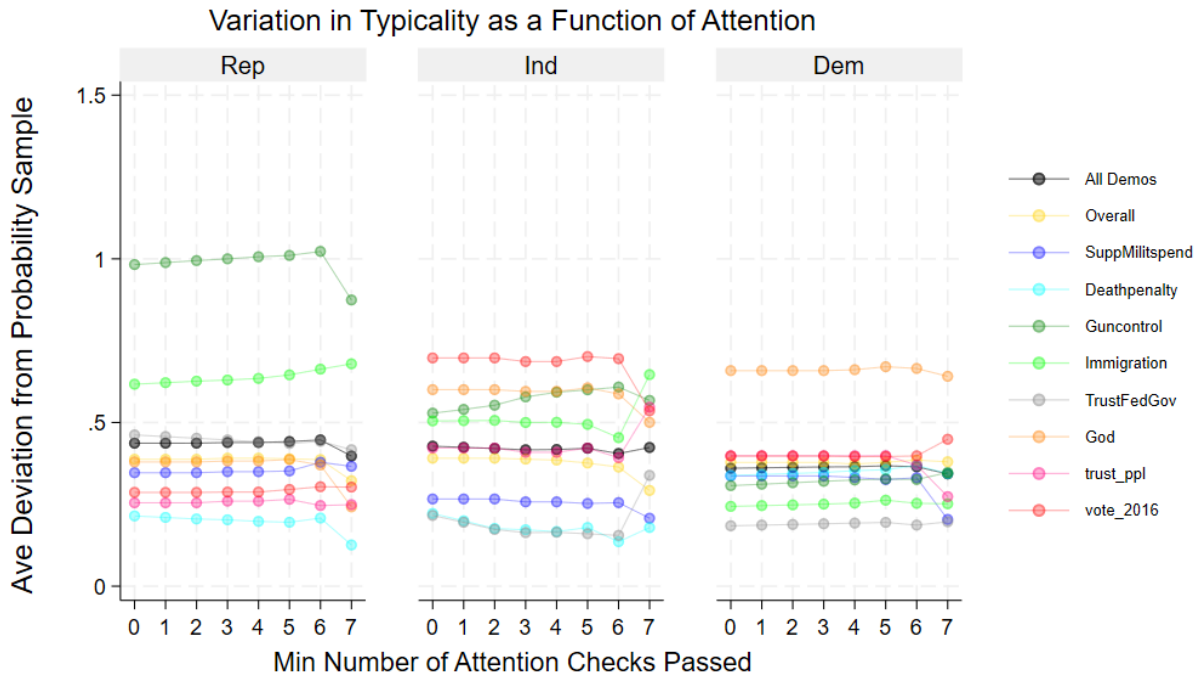
SI figure 9-7. *Representativeness for each Characteristic broken down by political party affiliation and level of attention check passed, for the sample Prolific.*

Prolific_NR:



SI figure 9-8. *Representativeness for each Characteristic broken down by political party affiliation and level of attention check passed, for the sample Prolific NR.*

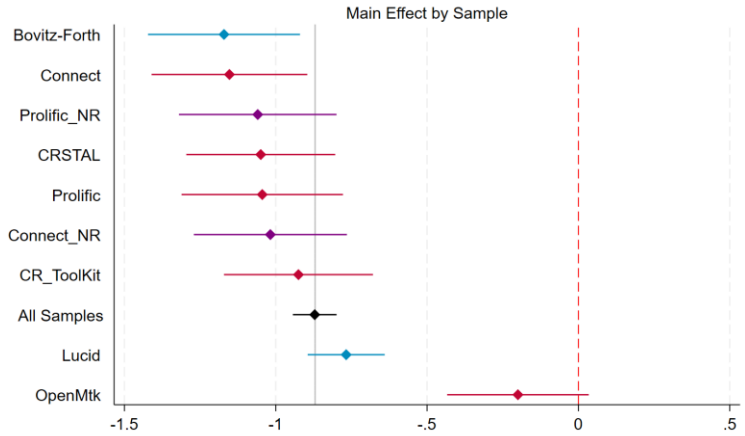
Connect_NR:



SI figure 9-9. *Representativeness for each Characteristic broken down by political party affiliation and level of attention check passed, for the sample Connect NR.*

SI- 10 Helping the Poor Policy Frame - Main effect

Using ordinal logistic regression, we replicate the expected main effect collapsing across samples (*b*= -.883, *p*< .001, *95%CI* [-.956, -.811]). Though effect sizes vary in magnitude and precision, we see the effect also essentially replicate across each sample individually, with the largest effect being Bovitz-Forthright (*b*= -1.17, *p*< .001) and the smallest being Open Mturk (*b*= -0.2, *p*= .094), which is the only sample that produces a non-significant effect at the .05 level.



SI Figure 10-1. *Coefficient plot for main effect of policy frame on support, broken down by sample. Black coefficient represents main effect collapsing across samples; colored coefficients indicate the three levels of representativeness as used in the main manuscript (blue = high representative, purple = mid representative, red = not representative).*

SI- 11 Loss Aversion Study Details

For the loss aversion paradigm, we presented participants with one of the two following frames:
Probabilistic gains frame:

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.
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.

Probabilistic losses frame

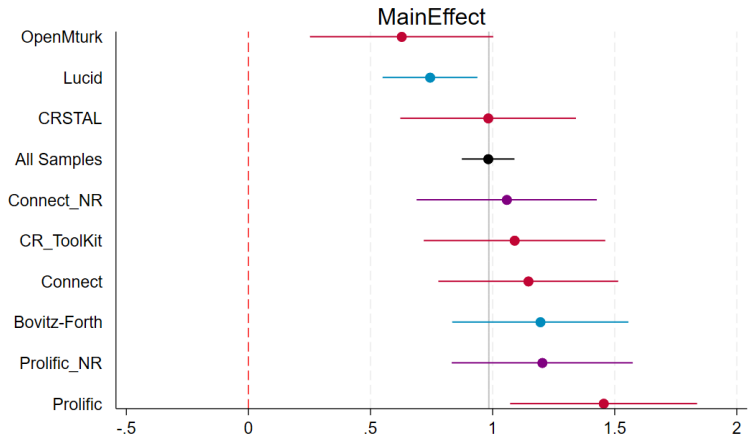
Imagine that your country is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed.
If Program A is adopted, 400 people will die.
If Program B is adopted, there is one-third probability that nobody will die, and two-third probability that 600 people will die.

Participants then chose one of the following options:

Based on the above information, what program would you choose?
Program A
Program B

SI- 12 Loss Aversion Main Effects by Sample

Using a logistic regression, we predict the probability of selecting the probabilistic option over the concrete option as a function of frame (loss vs gain). We successfully replicate previous work, showing that framing the effect of a policy in terms of losses produces an increase in selecting the probabilistic option. This is true overall (b= .988, p< .001, 95%CI[.88, 1.1]), and in each sample, with the largest effect being for Prolific (b=1.46, p< .001) and the smallest effect being for Open Mturk (b= .627, p= .001).



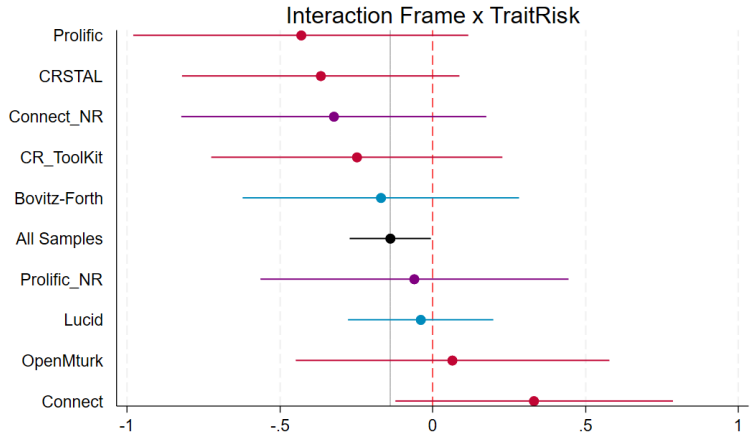
SI Figure 12-1. *Coefficient plot showing the effect of loss frame on policy selection, for each of the samples in color (blue = high representative, purple = mid representative, red = not representative) and the overall sample coefficient in black.*

SI- 13 Loss Aversion Interaction Effects by Sample

Looking at heterogeneous effects with this paradigm, previous work has identified an interaction between frame and risk preferences (Kam & Simas, 2010), showing that those higher on trait risk show a smaller change between policy frames. We attempt to replicate this here using risk items from the World Values Survey, American National Election Study, and a standard self-report risk item from political science (Kam, 2012)[27]. The alpha for this aggregate was low (a = .359), so we show results using the aggregate, and all three individually below.

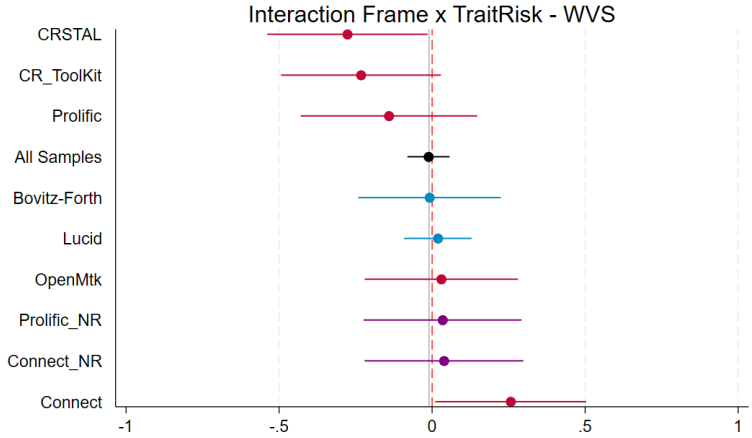
[27] Kam, C. D. (2012). Risk attitudes and political participation. *American Journal of Political Science*, *56*(4), 817-836.

The interaction between trait risk and loss frame is not statistically significant in any of the individual samples, but does show a barely significant effect when the samples are combined (p = .04). This indicated that those higher on trait risk were more likely to choose the probabilistic option overall (M= .433, SE=.01) compared to those lower in trait risk (M= .405, SE=.01); and those with higher trait risk showed a smaller difference in choosing the probabilistic option between frames (loss frame, *M*=0.54, *SE*=.013; gains frame, *M*=.327, *SE*=.012), compared to those lower on trait risk (loss frame, *M*=0.528, *SE*=.013; gains frame, *M*=.274, *SE*=.012). Further, there is some evidence of variation across samples (heterogeneity chi-squared between *p* = 0.003).

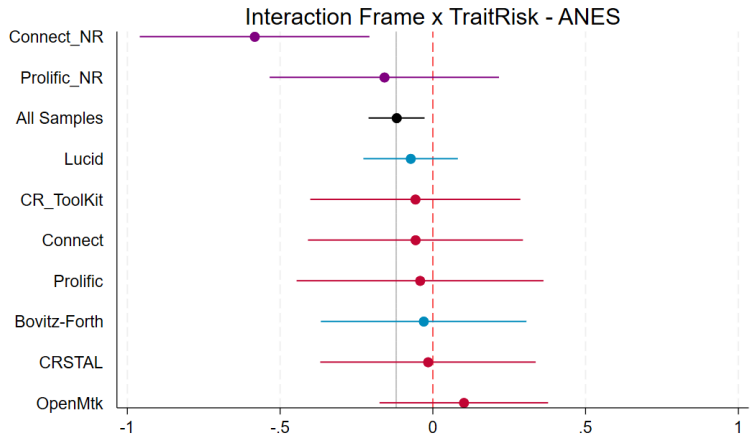


SI Figure 13-1. *Interaction coefficient between loss frame and trait risk aggregate. Each of the samples is in color (blue = high representative, purple = mid representative, red = not representative) and the overall sample coefficient is in black.*

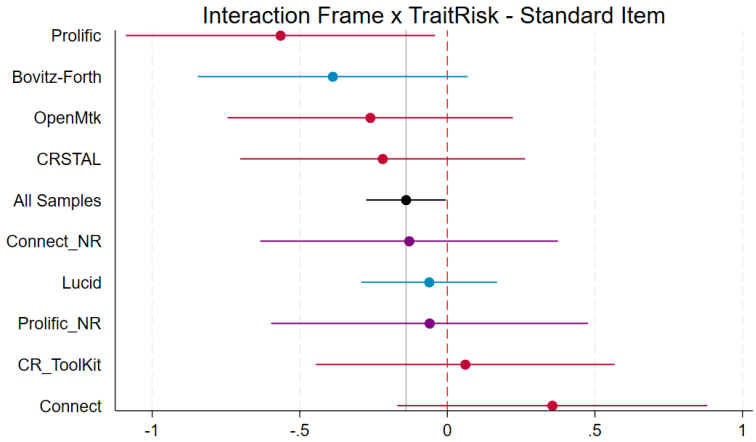
Below we also show the interactions for the subcomponent versions of the different risk measures.



SI Figure 13-2. *Interaction coefficient between loss frame and trait risk taken from the World Values Survey.*



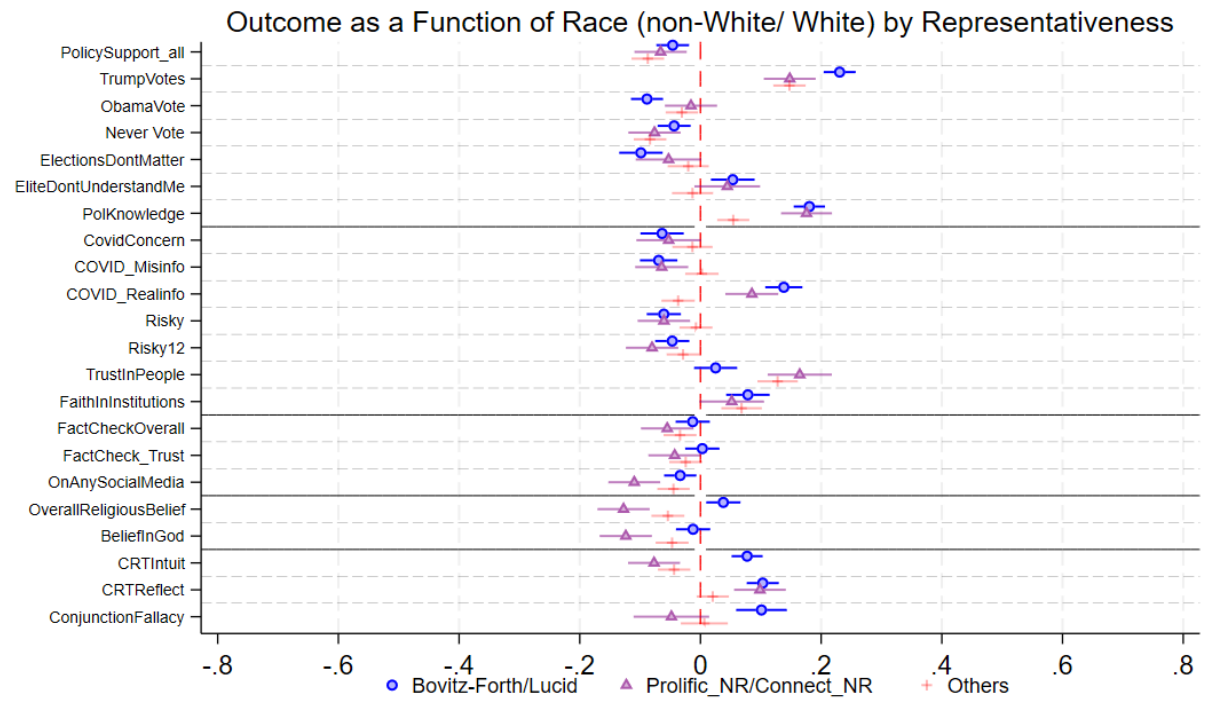
SI Figure 13-3. *Interaction coefficient between loss frame and trait risk taken from the American National Election Survey.*



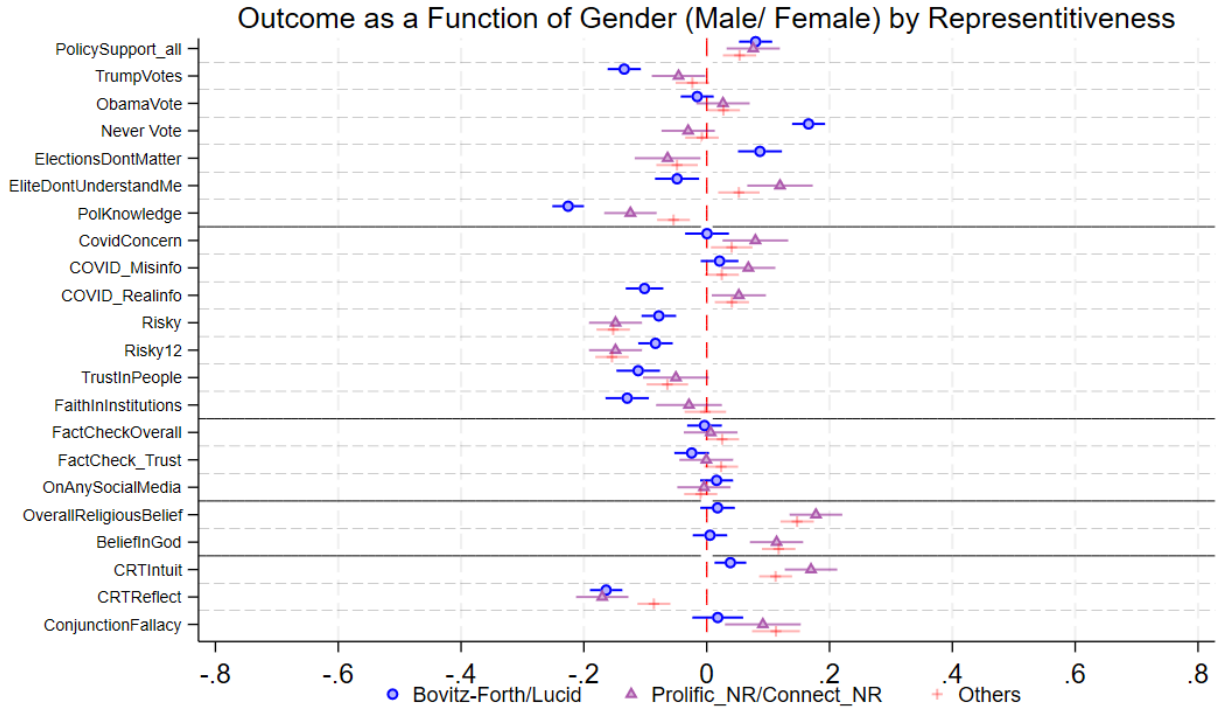
SI Figure 13-4. *Interaction coefficient between loss frame and trait risk represented by a standard risk item from political science.*

SI- 14 Exhaustive Comparison Analysis

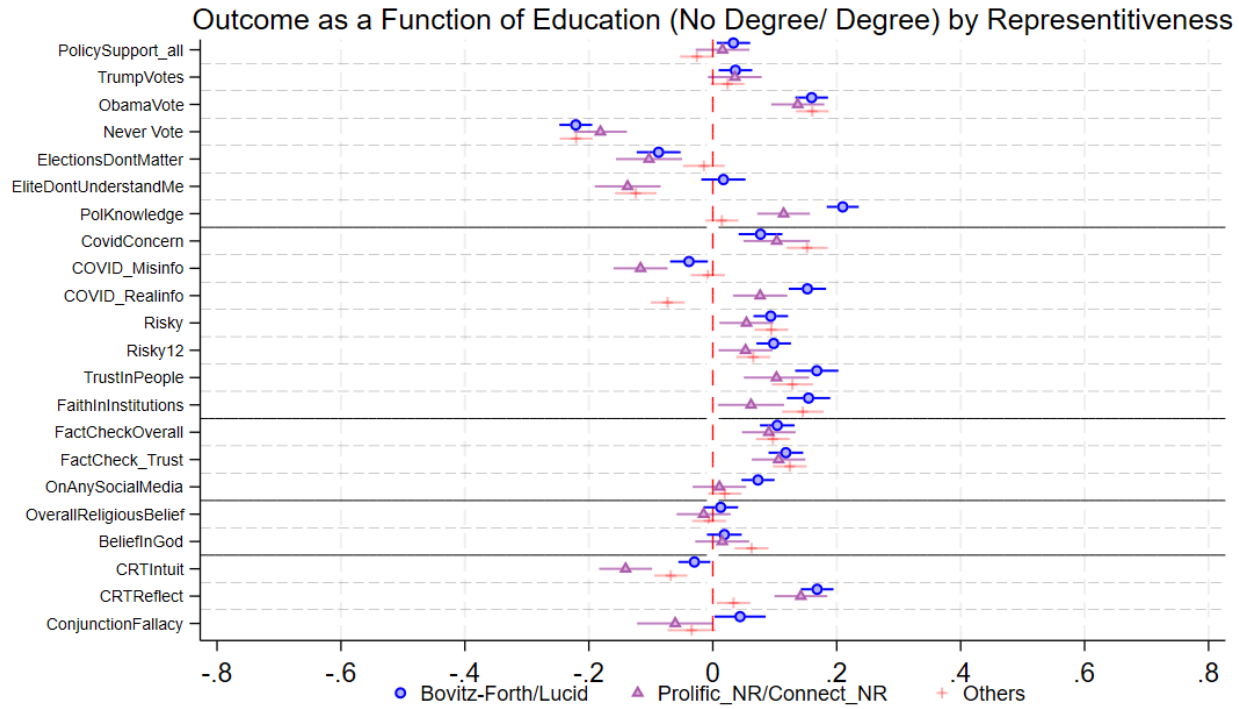
Here we provide a full comparison of six key demographics (race, age, education, gender, income and political affiliation) across 22 outcomes (attitudes, beliefs, knowledge and cognitive measure) for the three levels of representativeness groupings. The goal is to provide the reader with an exhaustive and unbiased comparison for any and all comparisons one may be interested in. Though, by design, there are no hypotheses being tested here, we still encourage the use of caution in interpreting any one result, given the very large number of comparisons being made. Across all relationships, we see notable discrepancy between the more versus less representative samples, suggesting heterogeneity, and the need for more representative samples depends on the presence of social/political content, though further work is still needed.



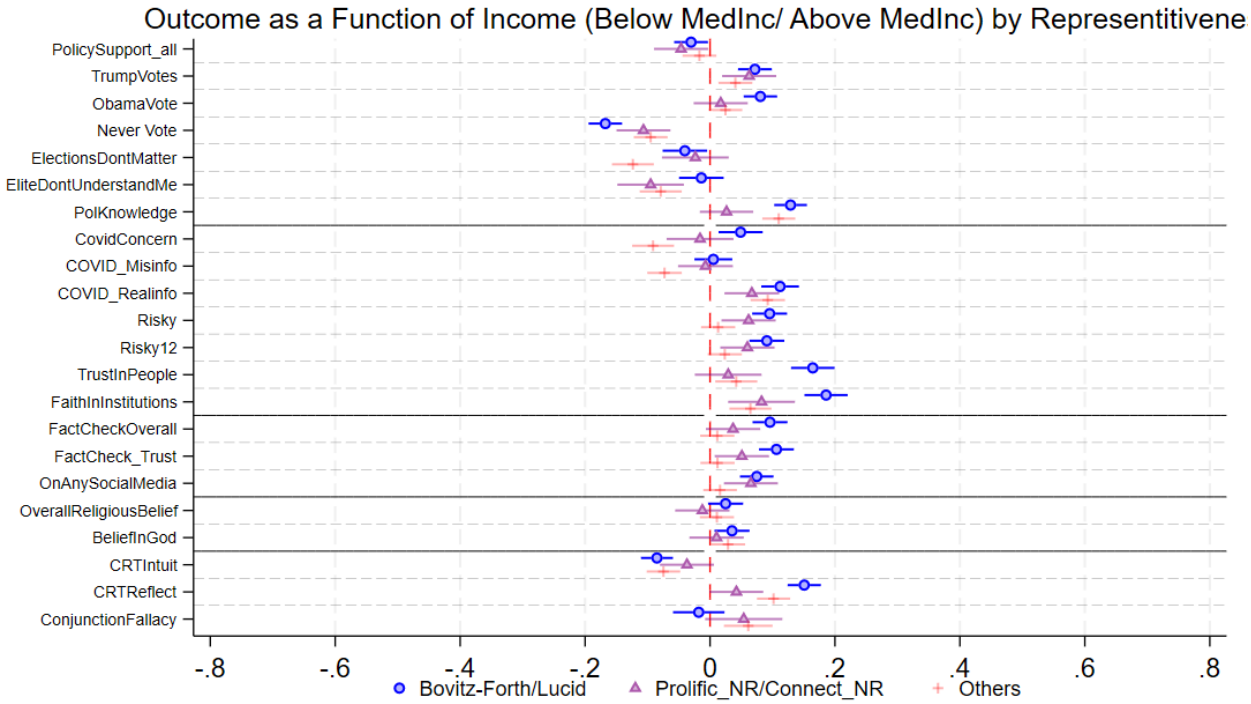
SI Figure 14-1. *Exhaustive unbiased comparison of 22 outcomes as a function of participant race and sample representativeness.*



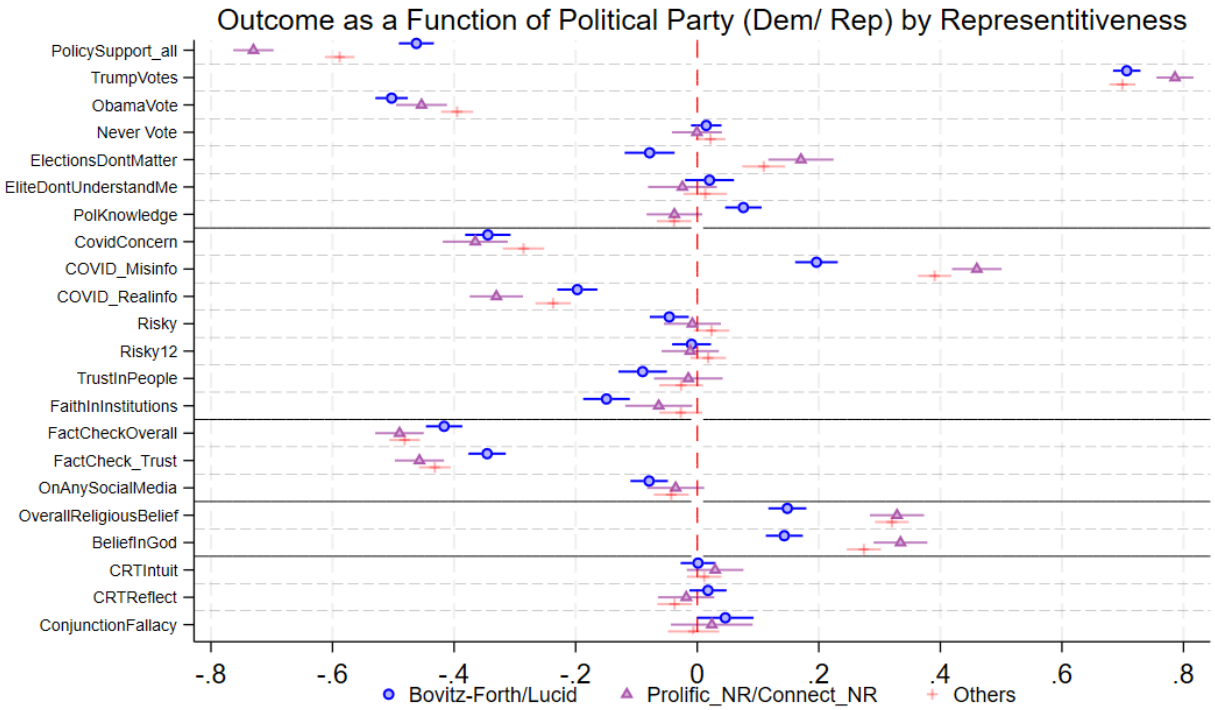
SI Figure 14-2. *Exhaustive unbiased comparison of 22 outcomes as a function of participant gender and sample representativeness.*



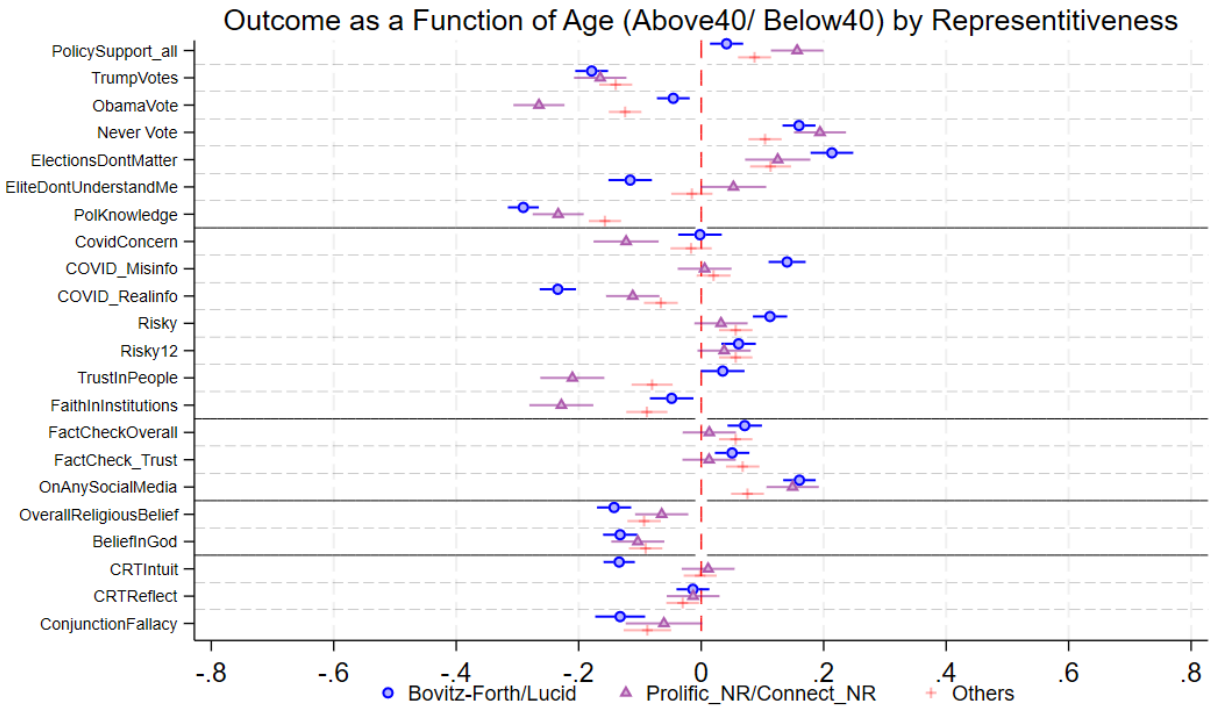
SI Figure 14-3. *Exhaustive unbiased comparison of 22 outcomes as a function of participant education and sample representativeness.*



SI Figure 14-4. *Exhaustive unbiased comparison of 22 outcomes as a function of participant income and sample representativeness.*



SI Figure 14-5. *Exhaustive unbiased comparison of 22 outcomes as a function of participant political party affiliation and sample representativeness.*

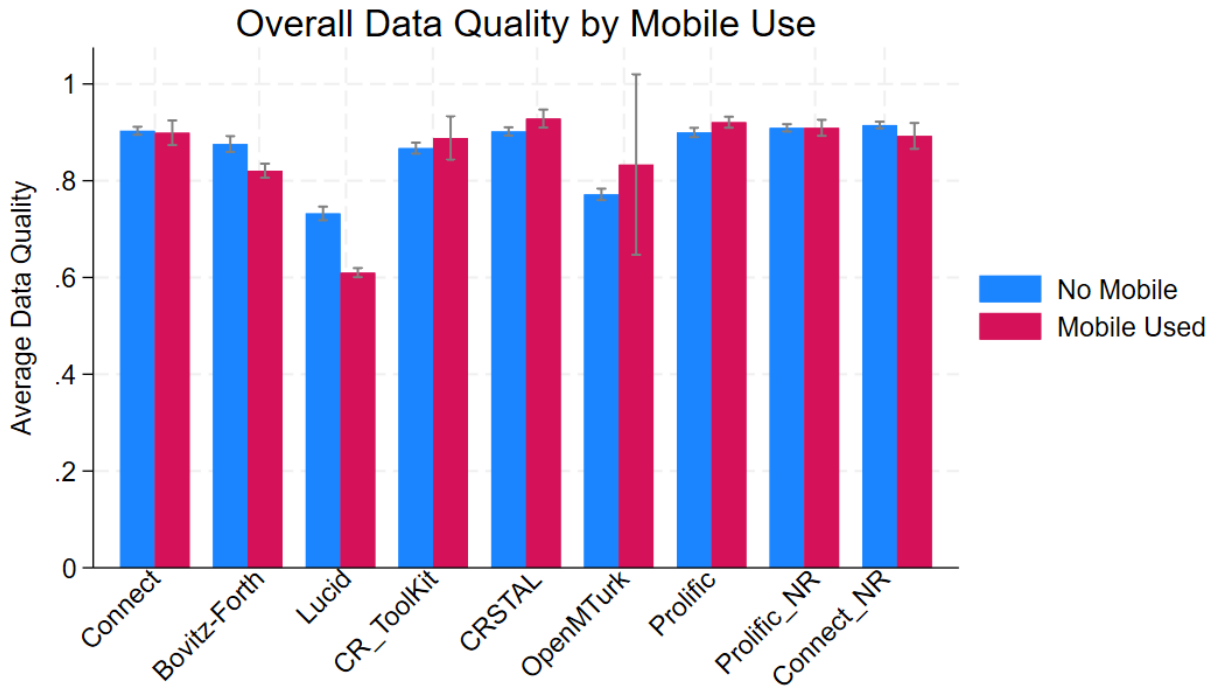


SI Figure 14-6. *Exhaustive unbiased comparison of 22 outcomes as a function of participant age and sample representativeness.*

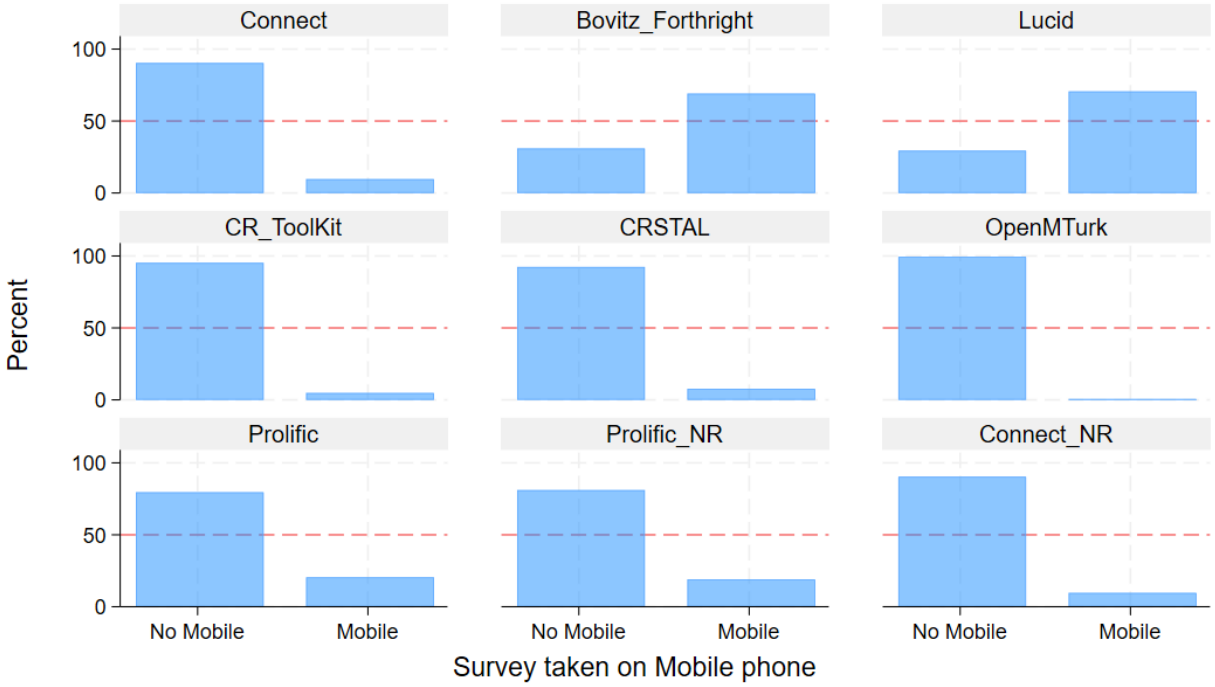
SI- 15 Mobile Phone Use and Response Quality

Here we breakdown response quality by mobile phone use for each sample. We see that though almost all samples show lower response quality as a function of taking this study on a mobile phone, the samples themselves initially show lower response quality due to greater phone usage.

SI Figure 15-1. *Response quality as a function of mobile phone use across samples. Note that for some samples, the underlying number of mobile users is quite low, see SI## below.*



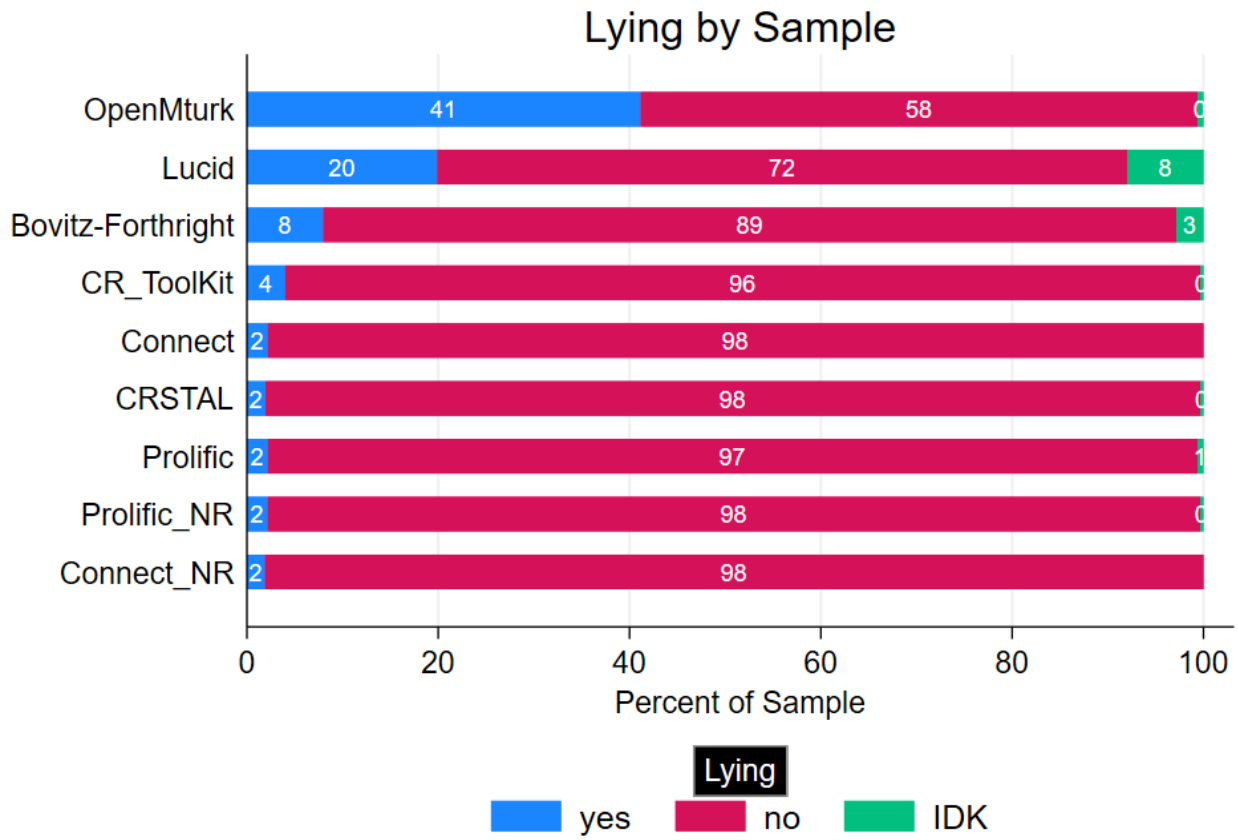
SI Figure 15-2. *Mobile phone use across samples.*



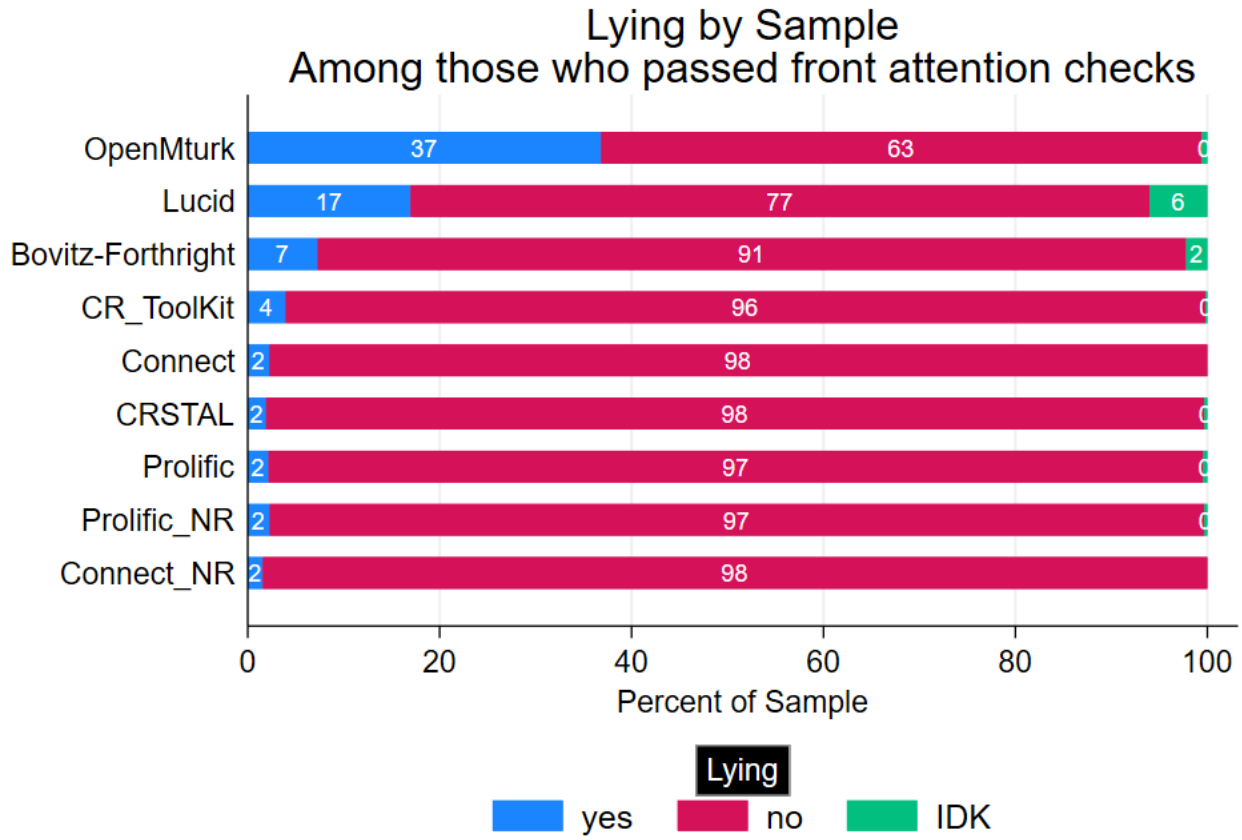
SI- 16 Lying and Honesty Across Samples

Here we break down the specific level of reported lying/ honesty across all samples. Note that front end attention checks interestingly appear to do little to curb lying, even among the top levels of attention (*SI16-3*).

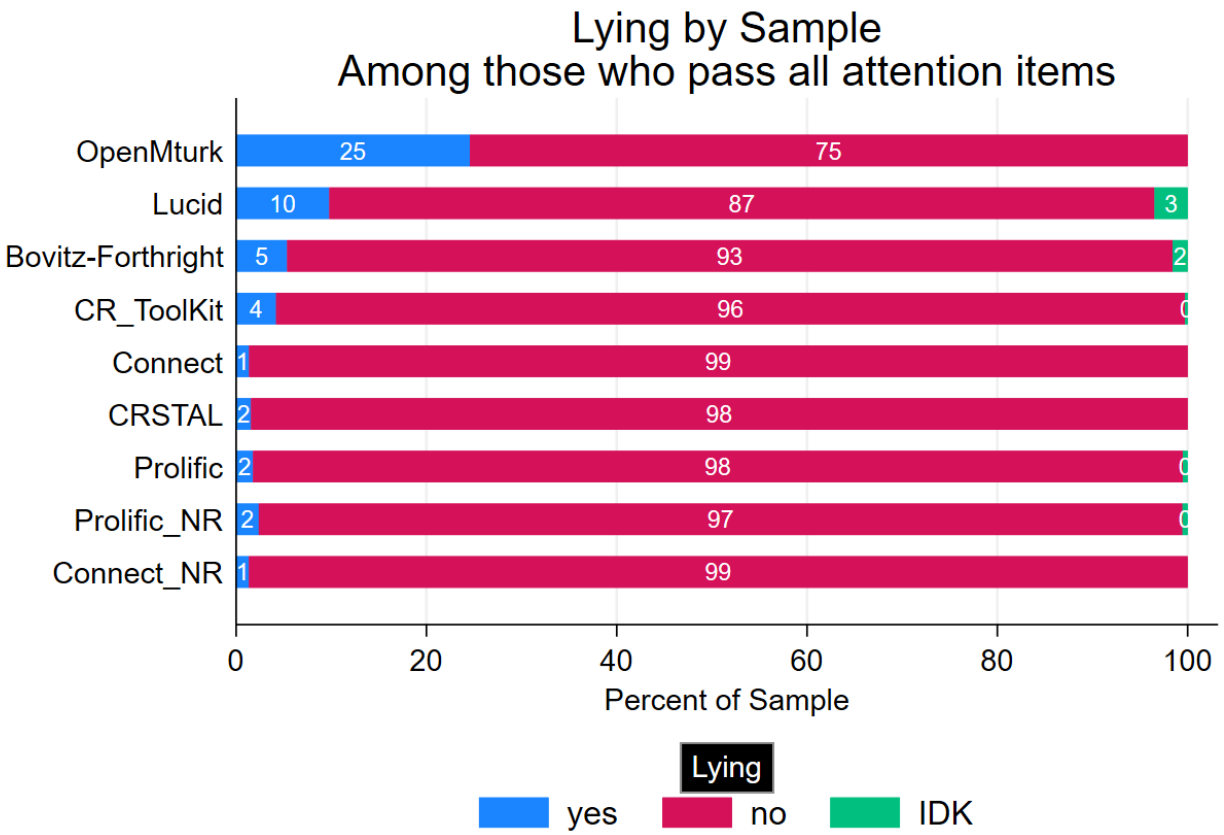
SI Figure 16-1. Self reported lying across samples.



SI Figure 16-2. *Self reported lying across samples among those who pass front end attention filters.*



SI Figure 16-3. *Self reported lying across samples among those who pass all attention filters.*



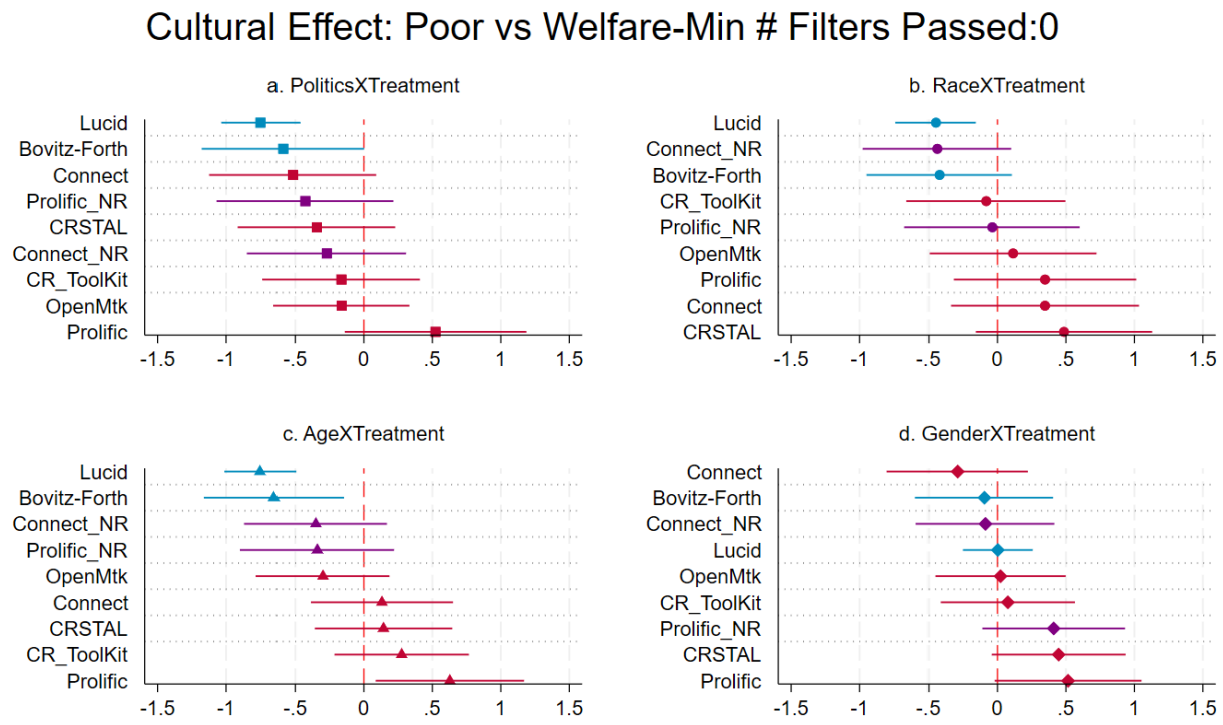
SI Table 16-1. Depicts multiple models using ordinal logistic regression predicting the three levels of honesty/lying for the three categories of representativeness (outcome scored such that higher scores indicates lying). Model one uses maximum representativeness samples, model two uses maximum representativeness and constraints to those who passed the first two attention checks. Model three uses mid level representativeness, model four uses mid level representativeness and constraints to those who passed the first two attention checks. Model five uses non-representative samples, and model six uses non-representative and constraints to those who passed the first two attention checks. Each model includes platform dummies due to the different samples that make up the representativeness categories.

	1	2	3	4	5	6
VARIABLES	Max_NR	Max_NR AttCk	Mid_NR	Mid_NR AttCk	Not_NR	Not_NR AttCk
Age	-0.654***	-0.716***	-0.911***	-0.850***	-0.220**	-0.226**
	(0.0429)	(0.0523)	(0.229)	(0.240)	(0.0813)	(0.0868)
Gender	-0.193*	-0.117	-0.117	0.0426	-0.349**	-0.300*
	(0.0860)	(0.104)	(0.326)	(0.341)	(0.113)	(0.120)
Income	0.0348	0.0469	-0.0771	-0.111	-0.377***	-0.378***
	(0.0415)	(0.0527)	(0.174)	(0.184)	(0.0730)	(0.0775)
College degree	-0.0949	-0.0790	-0.143	-0.294	0.439**	0.423**
	(0.0915)	(0.110)	(0.349)	(0.363)	(0.153)	(0.161)
Urbanr ural	-0.0719+	-0.0989*	-0.440*	-0.469*	-0.0223	-0.0109
	(0.0383)	(0.0484)	(0.195)	(0.207)	(0.0559)	(0.0602)
Political Ideology	-0.0638	-0.00675	0.144	0.0426	0.308***	0.330***
	(0.0429)	(0.0517)	(0.174)	(0.187)	(0.0529)	(0.0569)
EthnicRace-Black	0.261*	0.327*	1.196**	1.494***	0.712***	0.826***
	(0.110)	(0.141)	(0.384)	(0.412)	(0.201)	(0.209)
EthnicRace-Latino	0.343*	0.159	0.882+	1.242*	0.217	0.245
	(0.170)	(0.213)	(0.514)	(0.539)	(0.320)	(0.340)
EthnicRace-Asian	0.348+	0.208	1.070*	1.398*	0.564*	0.618**
	(0.199)	(0.248)	(0.539)	(0.564)	(0.227)	(0.232)
EthnicRace-Other	0.166	0.0919	-11.82	-12.60	0.0942	-0.541
	(0.283)	(0.349)	(632.1)	(1,050)	(0.793)	(1.069)
EthnicRace-Mix	-0.301	-0.253	0.472	0.794	0.972***	1.057***
	(0.212)	(0.246)	(0.781)	(0.800)	(0.236)	(0.248)
Lucid	1.269***	1.252***				
	(0.115)	(0.130)				
Connect_NR			-0.426	-0.664+		
			(0.335)	(0.360)		
CR_ToolKit					0.594*	0.521+
					(0.268)	(0.272)
CRSTAL					-0.0284	-0.0759
					(0.304)	(0.307)
OpenMturk					3.128***	2.937***
					(0.232)	(0.235)
Prolific					0.248	0.131
					(0.291)	(0.298)
Observations	4016	3238	1985	1907	4976	4750
Pseudo R2	.092	.096	.125	.146	.315	.293

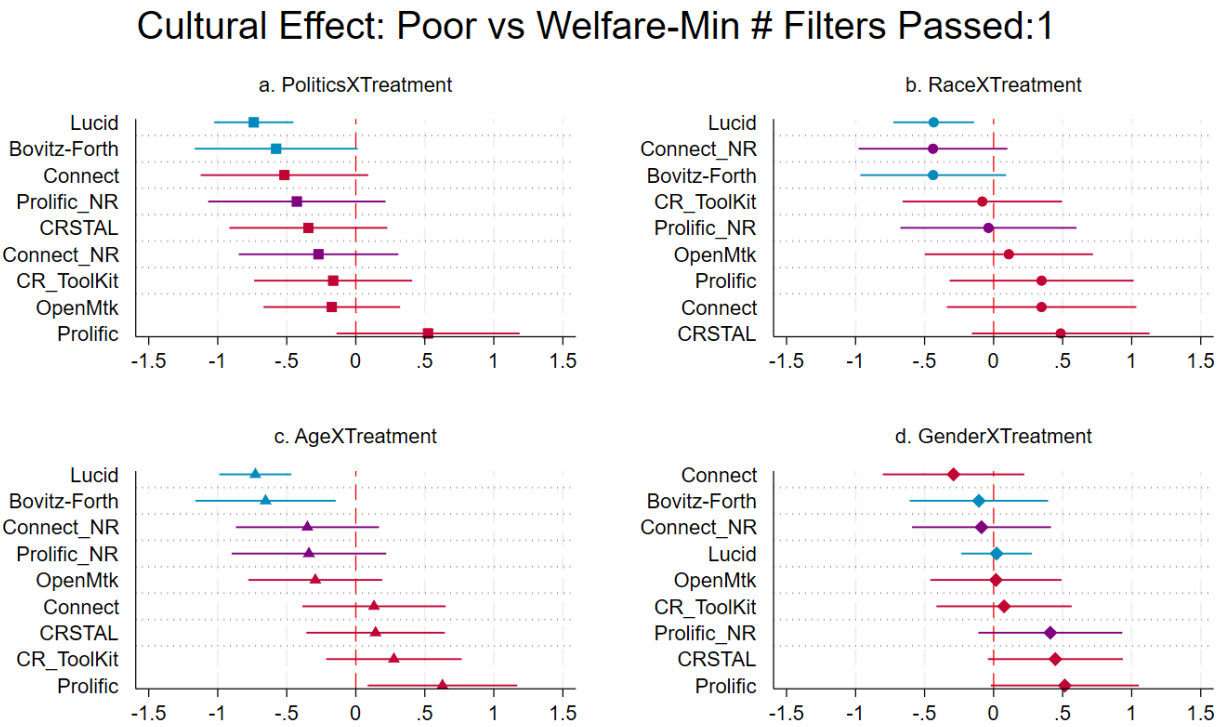
Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

SI- 17 Effect of filtering on treatment effects *As above with response quality and representativeness, here we show the effect of the constraining to different levels of attention filtering on the above displayed interactions between key demographics and treatment effect. Overall, we see that the ranks stay stable up till roughly the mid-level of filtering.*

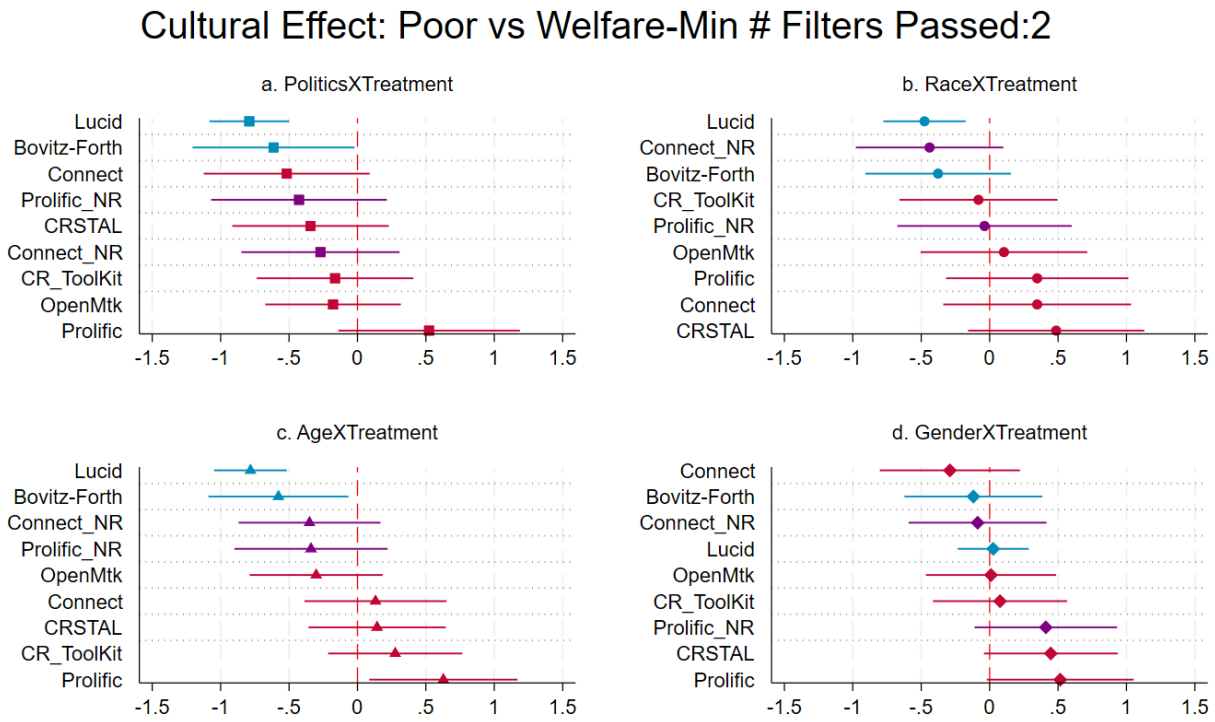
SI Figure 17-1. Zero levels of attention filtering (no filtering) on the displayed interactions between key demographics and treatment effect.



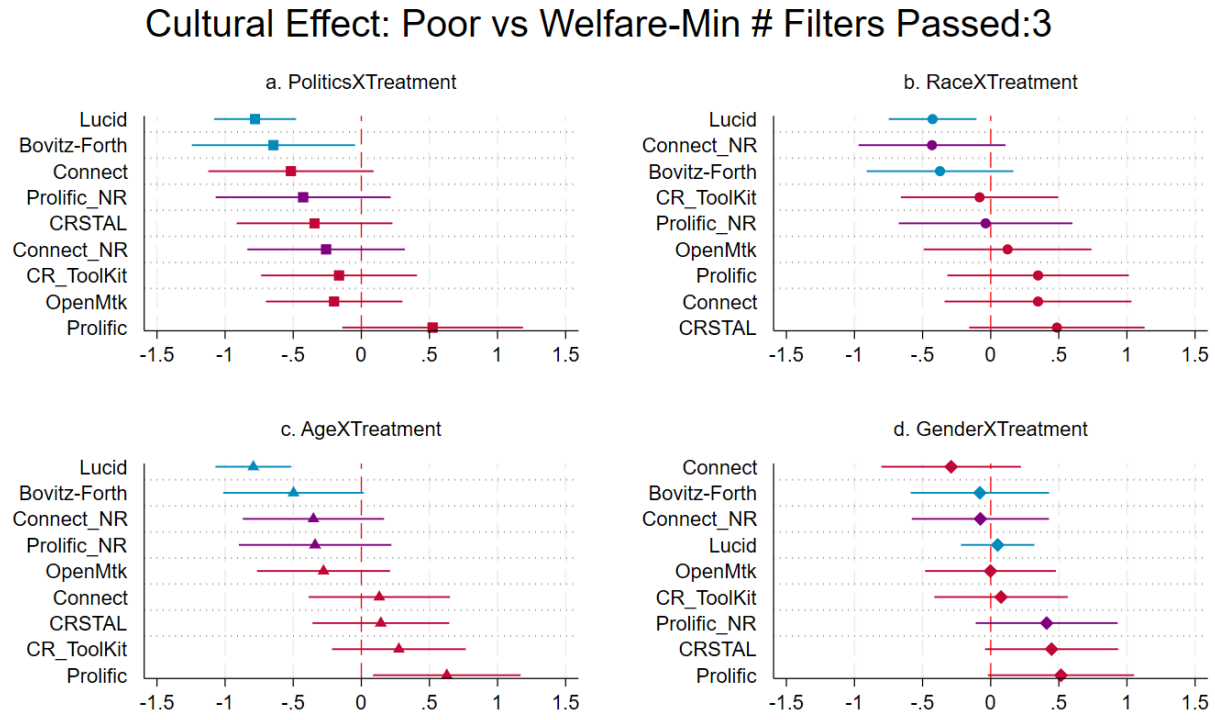
SI Figure 17-2. Filtering on one level of attention for the interactions between key demographics and treatment effect.



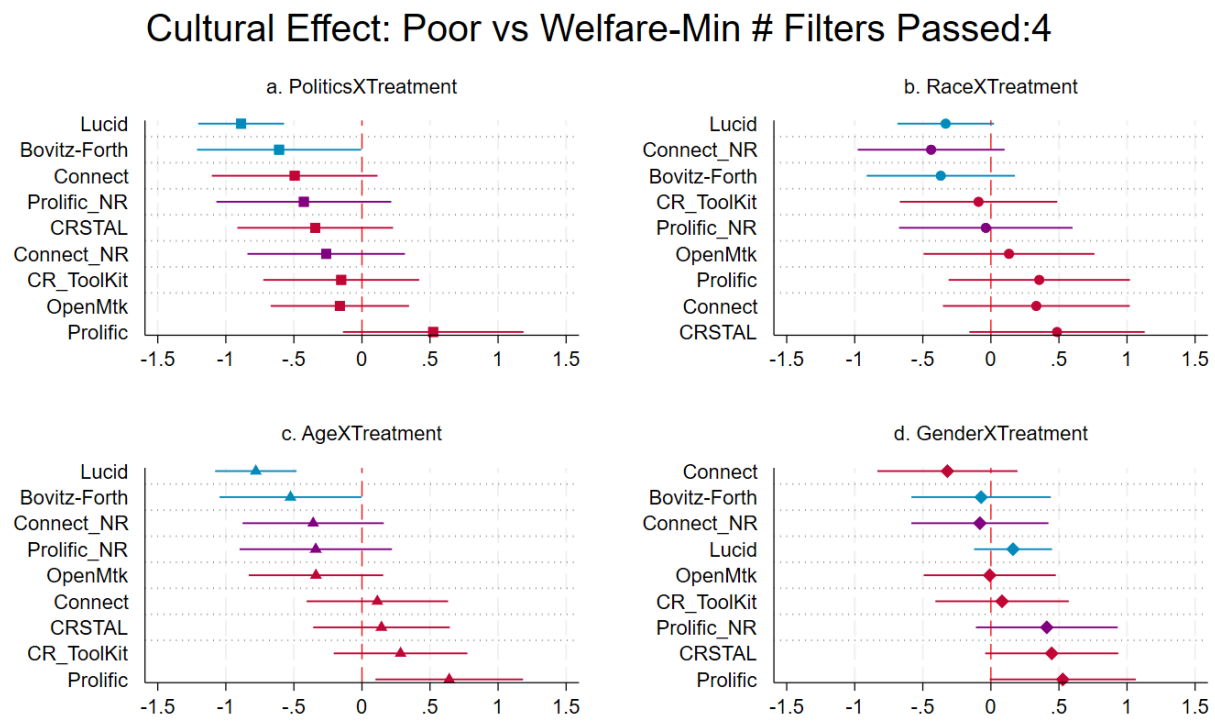
SI Figure 17-3. Filtering on two levels of attention for the interactions between key demographics and treatment effect.



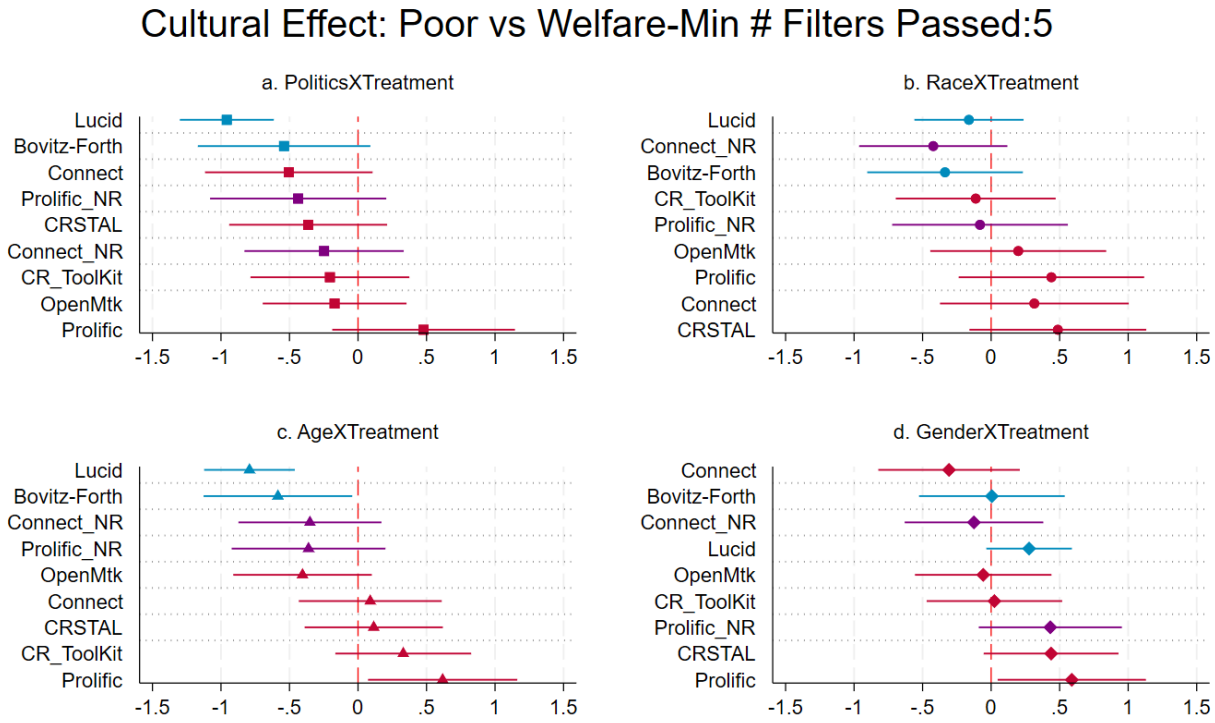
SI Figure 17-4. Filtering on three levels of attention for the interactions between key demographics and treatment effect.



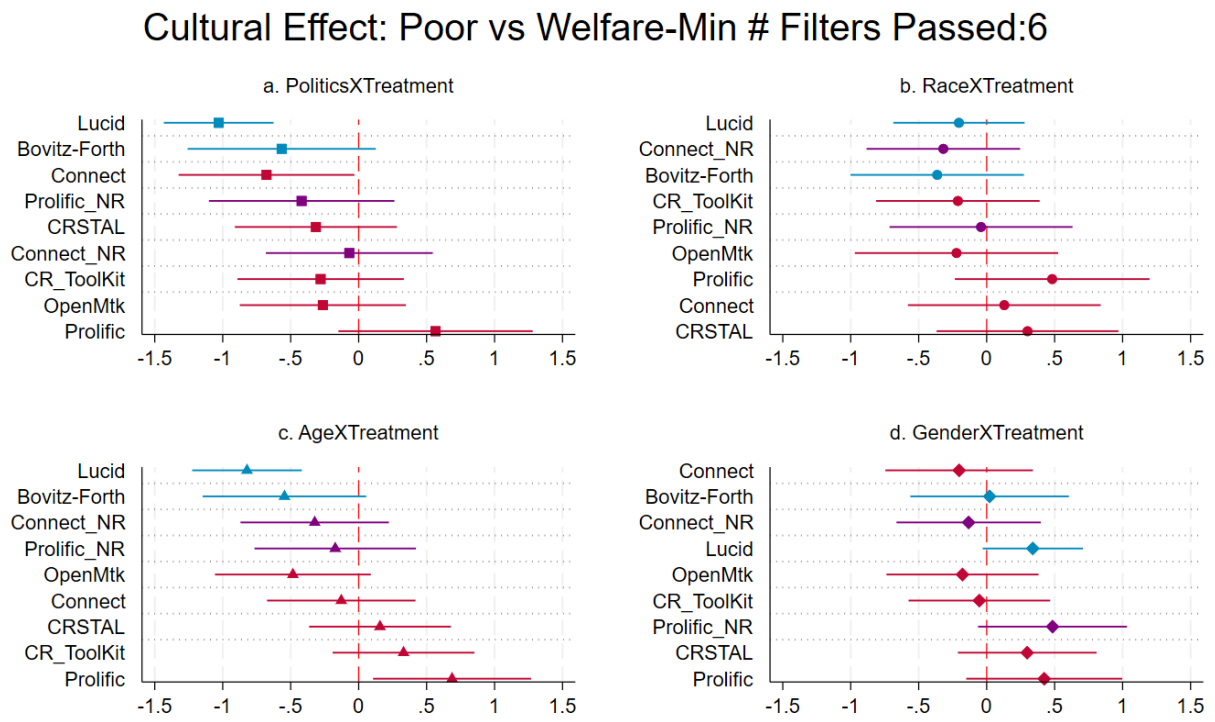
SI Figure 17-5. Filtering on four levels of attention for the interactions between key demographics and treatment effect.



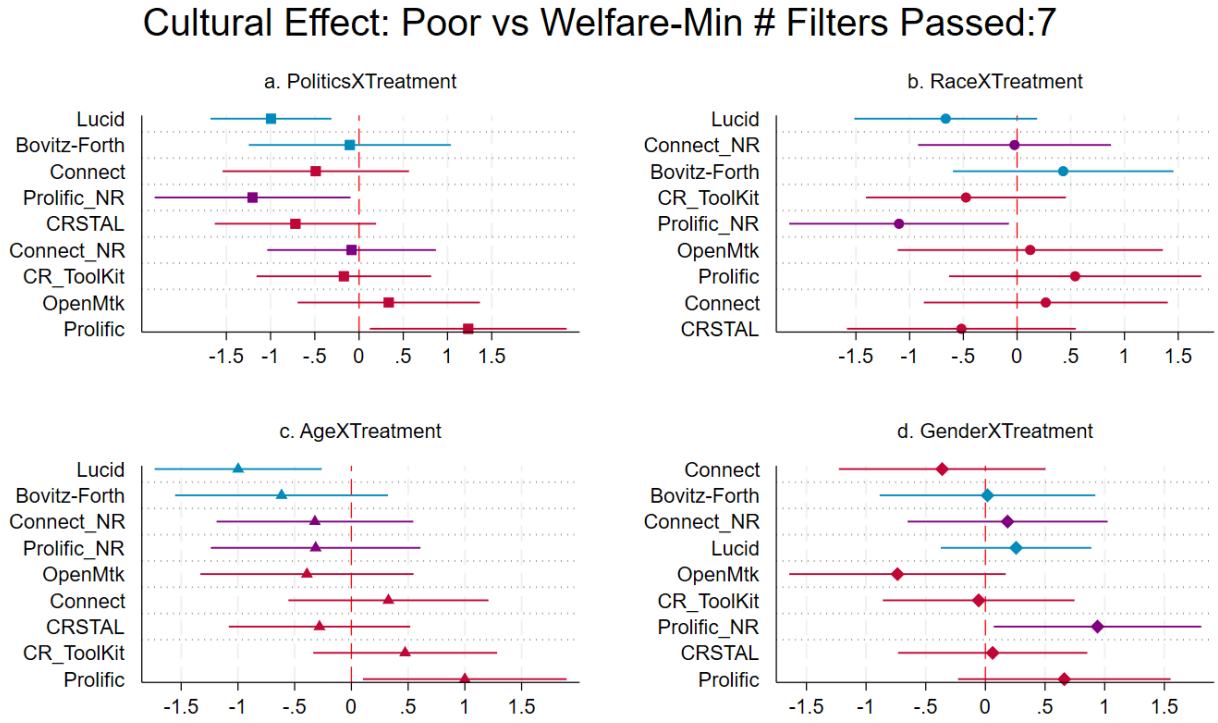
SI Figure 17-6. Filtering on five levels of attention for the interactions between key demographics and treatment effect.



SI Figure 17-7. Filtering on six levels of attention for the interactions between key demographics and treatment effect.



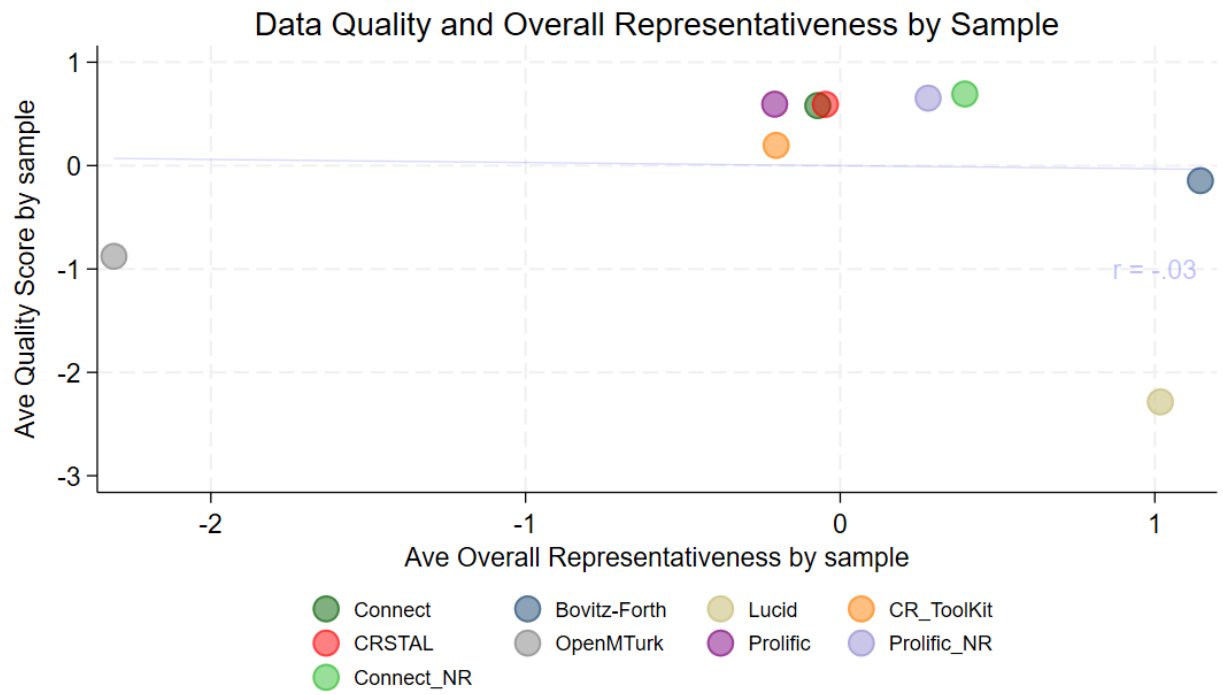
SI Figure 17-8. Filtering on all levels of attention for the interactions between key demographics and treatment effect.



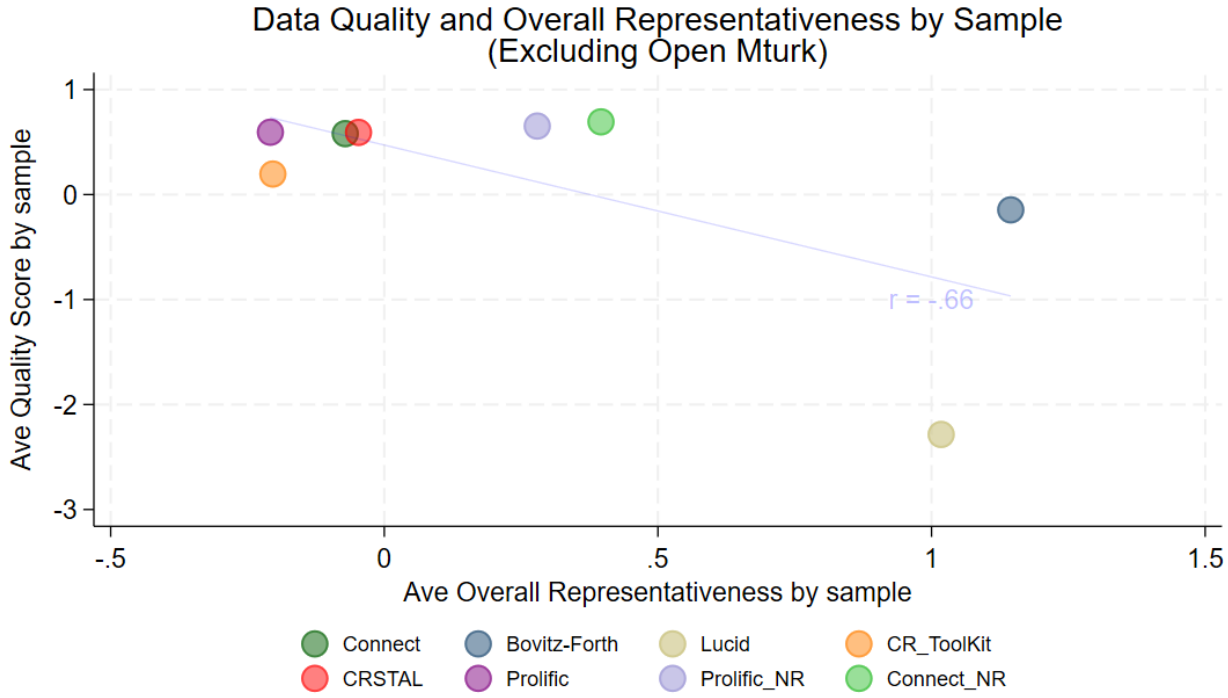
SI- 18 The Relationship Between Overall Representativeness and Response Quality

Here we calculate a measure of overall representativeness, averaging both Typicality and Demographic representativeness into one value, as well as a single value for response quality for each sample. We show that there is a strong negative relationship between these two dimensions. Further, this relationship is hidden by the notably low performing OpenMturk sample, which is a notable outlier compared to the others on both dimensions. With OpenMturk in the model, and due to the small sample of platforms we compare, the negative relationship disappears.

SI Figure 18-1. *Response quality by overall representativeness.*



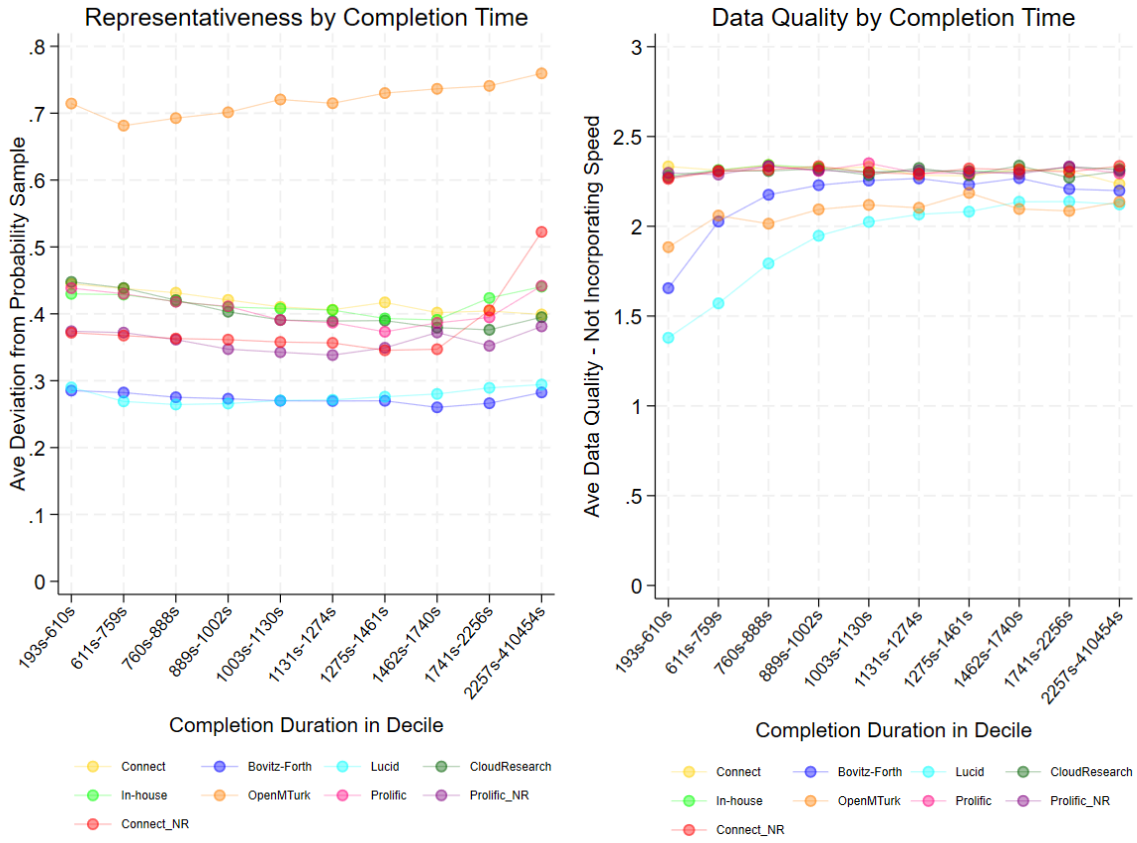
SI Figure 18-2. *Response quality by overall representativeness removing the Open Mturk sample.*

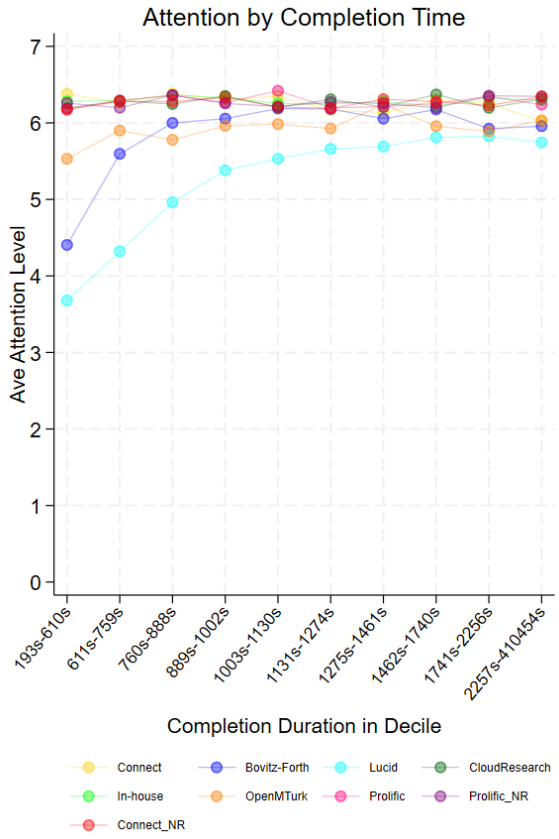


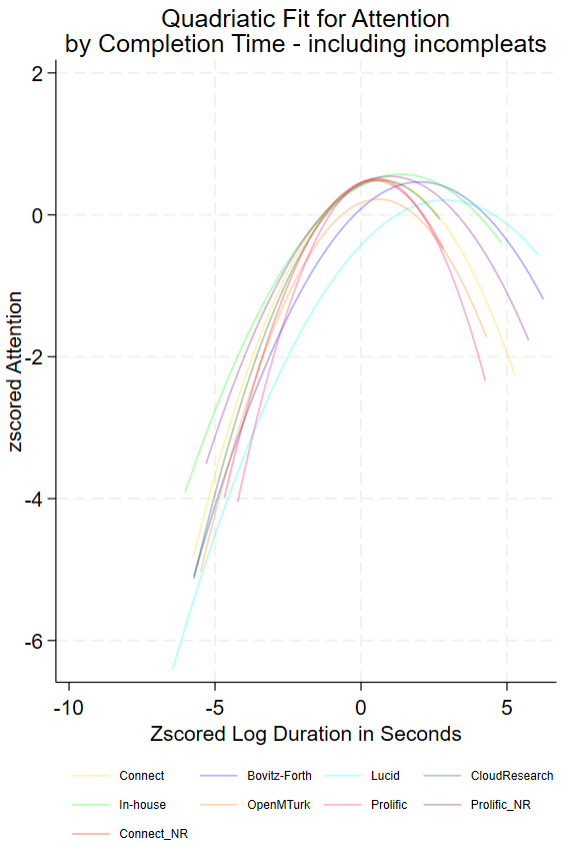
SI- 19 Relationship between Duration, Representativeness, Response Quality and Attention

Here we first look at the relationship between completion time (broken into decile) and demographic representativeness for each sample. We see little variation across samples with a broad loss of representativeness at the slower ranges of completion times (note, only those who

completed the full study are included in this analysis). Looking at overall response quality (removing speeding from the aggregate quality measure), we see faster duration times are associated with lower response quality. This was mainly the case for the three lower response quality samples, with little variation among those who were at the ceiling for quality in the main paper. This result is the same when looking at attention levels specifically. However, when modeling attention data by log duration times (including those who do not finish) with a quadratic fit, we see an inverted-U shape, indicating that those with very short or very long durations are lower on attention.



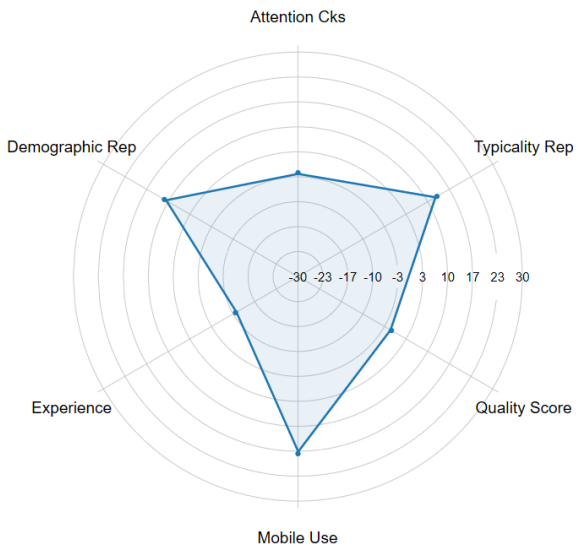




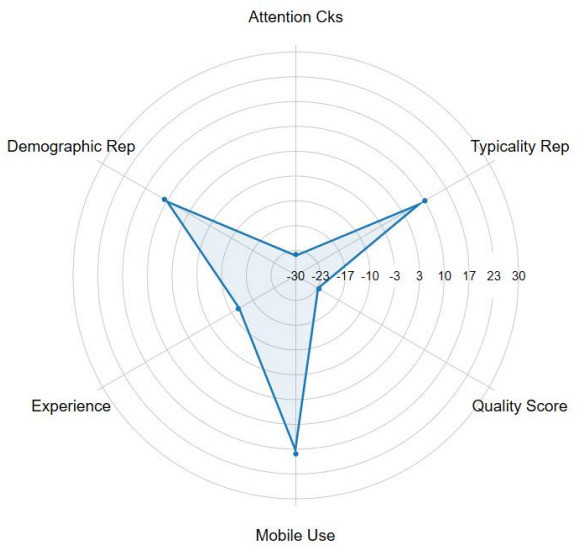
SI- 20 Sample Profile

Here we present radar plots that represent the overall profile of each sample on the dimensions of: demographic representativeness, typicality representativeness, overall response quality, attention, reported experience and mobile use. Note that higher scores indicate an increase in that dimension, including both versions of representativeness. All scales have been standardized and increased by an order of magnitude for plotting purposes.

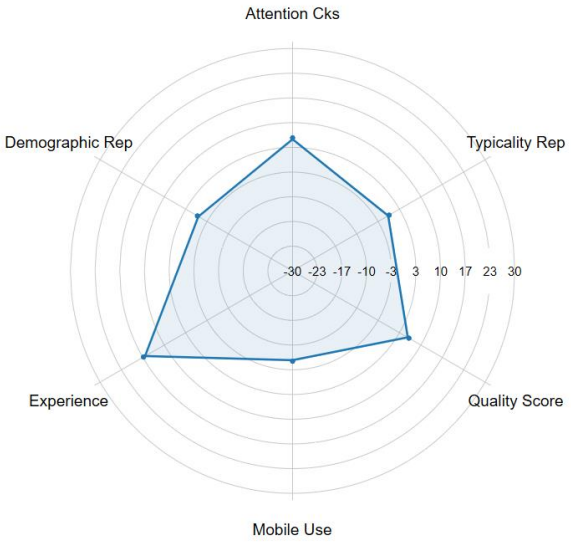
Bovitz-Forthright Profile



Lucid Profile

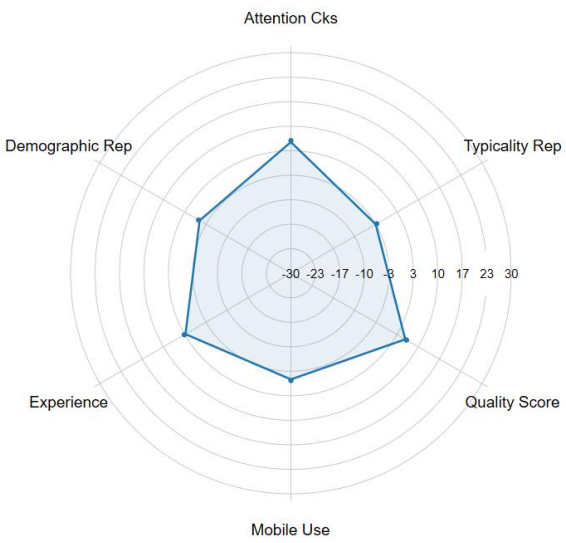


CRSTAL Profile

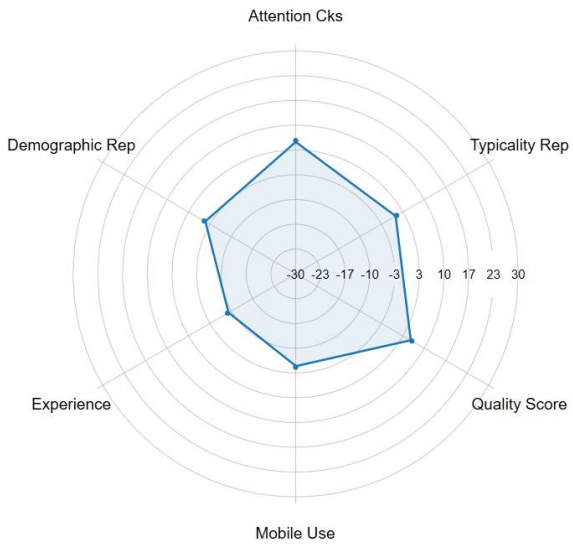


CR_ToolKit Profile

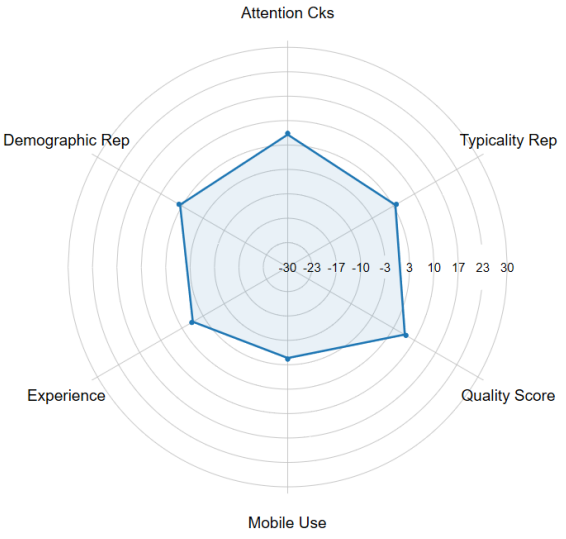
Prolific Profile



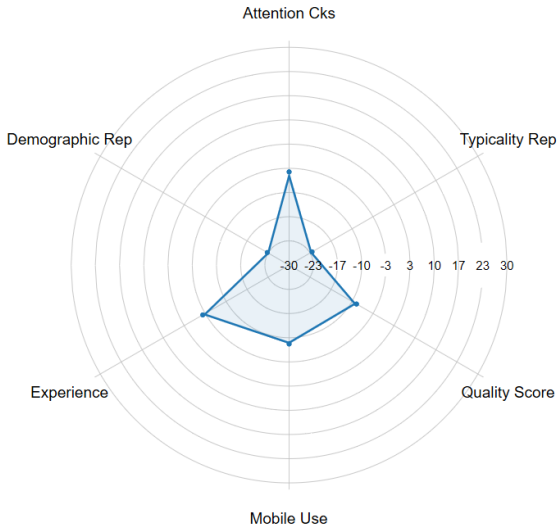
Connect Profile



Connect_NR Profile



OpenMTurk Profile



SI-21 Post stratification weighting and treatment effects

There is not a consensus on whether to include sample weights in experimental research (Franco et al. 2017). The inclusion of weights is helpful when the variables that are weighted moderate a treatment effect; the downside of weights is they reduce statistical power (Miratrix et al. 2018) and can introduce other types of biases (Winship and Radbill 1994, Franco et al. 2017). We explored the influence of adding demographic-based weights on treatment effects for the five less-representative samples that did not use quota matching: Connect, Prolific, CRSTAL, CR ToolKit and OpenMturk (see SI-4 for a demographic comparison of the samples). Specifically, we add post stratification weights, per platform, that match the benchmark distribution of gender, age, race and political party affinity (one at a time). Moreover, we also explore the influence of weighting all four dimensions at once on moderating treatment effects using a raking routine introduced in Kolenikov (2019). Further, we merge all five samples (to maximize power), as well as the same set of samples removing Open Mturk, and apply all four dimensions of weights at once. We do this to explore whether weights can reproduce treatment effects / moderation effects in a situation with a very large sample size.[28]

First, we note that the (unweighted) overall main effect reported for the experiment on social entitlement spending slightly changes when focusing on the non-representative samples (-.883 vs -.855, collapsing across samples). Including post stratification weights, we find that the main effect for non-representative samples becomes stronger with the race, party, and combined weights (Table 20-1). Treatment effects per platform reveal that the (main) effect becomes significant for *Open Mturk* with all the weights but gender, which suggest that a key driver for the overall improvement is due to the specific gains made by *Open Mturk*.

Indeed, a closer look at the sample composition for *Open Mturk* reveals that 23% of the participants identified with the following profile: white, strong democrat, and aged between 32 and 48; the proportion of participants with that same profile was only between 4% and 11% for the other platforms. Thus, the atypical prevalence of this group, relative to the other samples and likely to the population in the U.S., makes this sample particularly susceptible to drastic changes by demographic-based weights.

Sample	**No weights** (n=5,079)	**Gender weight** (n=5,079)	**Age weight** (n=5,078)	**Race weight** (n=5,079)	**Party weight** (n=4,986)	***All/raked* weights** (n=4,985)
1. Connect	-1.153***	-1.160***	-1.298***	-1.225***	-1.243***	-1.409***
2. CloudResearch's Mturk Toolkit	-.925***	-.924***	-.755***	-.950***	-1.042***	-1.072***
3. Stanford CRSTAL Mturk panel	-1.049***	-1.037***	-1.176***	-.998***	-1.069***	-1.125***
4. Open Mechanical Turk	-.200^{+}	-.206^{+}	-.510***	-.346**	-.439***	-1.295***
5. Prolific	-1.045***	-1.044***	-.694***	-1.188***	-1.021***	-1.132***
All 9 samples	*-.883****	-	-	-	-	-
Non-representative (1+2+3+4+5)	**-.855*** **	**-.854*** **	**-.886*** **	**-.918*** **	**-.961*** **	**-1.202*** **
Non-representative *minus* (4)	**-1.041*** **	**-1.039*** **	**-.980*** **	**-1.083*** **	**-1.093*** **	**-1.177*** **

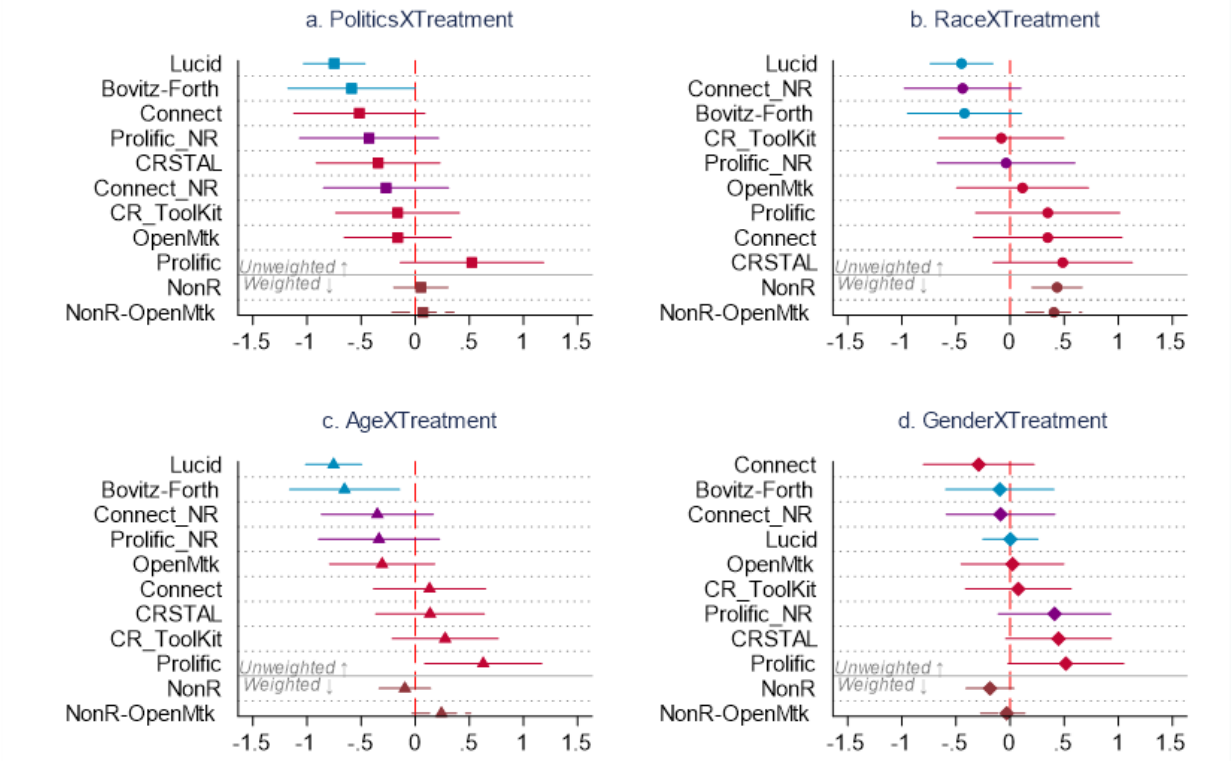
SI Table 21-1. *""Helping the poor" vs "Those on welfare" main effects, incl. demographic weights.Sample 4 and 6 had no records for the age group [81+] so their final category was [65+] when calculating their respective weights (results remain qualitatively similar if treated as missings); p<.001 ***; p<.01 **; p<.05 *; p<.1* $^{+}$

[28] We also tested the impact of weights on the other samples. They did not resuscitate CloudResearch or Prolific NR in terms of recovering the expected heterogeneous effects.

With regard to treatment effect heterogeneity across samples, recall that though we found (expected a la prior work; Green and Kern 20120) moderation by politics and age in the more representative samples (Bovitz and Lucid; see Figure 21-1), the non-representative samples did not show such moderation. Here with the addition of weights, individually or merged, we still do *not* find overall significant moderation effects of politics, age or gender (Table 21-2); there was no clear pattern for moderation with race (the direction gets reversed).

	Female (0) vs Male (1) x Treatment			**Under (0) vs Over 40 (1) x Treatment**			**Non-White (0) v White (1) x Treatment**			**Democrat (0) v Republican (1) x Treatment**		
	No weight	*Gender weight*	*Raked weights*	*No weight*	*Age weight*	*Raked weights*	*No weight*	*Race weight*	*Raked weights*	*No weight*	*Party weight*	*Raked weights*
1	-.290	.290	.357	.132	-.044	.362	.347	.382	-.452^{+}	-.517^{+}	-.506^{+}	-.314
2	.076	.076	-.512*	.276	.374	.059	-.081	-.007	.675**	-.163	-.255	-.245
3	.447^{+}	.447^{+}	.754**	.137	-.080	-.314	.486	.151	-.145	-.343	-.455	-.033
4	.022	.022	-.956***	-.307	-.626*	-1.460***	.114	.370	.618*	-.162	.003	-.122
5	.516^{+}	.516^{+}	-.052	.628*	.912**	-1.020**	.348	.542^{+}	1.627***	.524	.518	.759*
All	.114	-	-	-.345***	-	-	-.146^{+}	-	-	-.446***	-	-
NR	.200^{+}	.132	-.186	.043	.027	-.096	.225	.195	.434***	-.232^{+}	-.190	.052
N2	.188	.183	-.031	.266*	.266^{+}	.242^{+}	.227	.219	.404**	-.154	-.199	.072

SI Table 21-2. *Interaction between treatment and participant's gender (female vs male), age (above vs below 40 years), race (white vs not white), party affiliation (Republican vs Democrat), including demographic weights. Sample 2 and 4 had no records for the age group [81+] so their final category was [65+] when calculating their respective weights (results remain qualitatively similar if treated as missings). 1=Connect; 2=CloudResearch's Mturk Toolkit; 3=Stanford CRSTAL Mturk panel; 4=Open Mechanical Turk; 5=Prolific. p<.001 ***; p<.01 **; p<.05 *; p<.1 $^{+}$; p>.1*



SI Figure 21-1. *Interactions (as reported in the main section of the paper, unweighted) between treatment and participant's gender (female vs male), age (above vs below 40 years), race (white vs not white), party affiliation (Republican vs Democrat). Blue indicates the two most representative samples, purple indicates the two somewhat representative samples, red depicts the least representative samples. We now add dark red, depicting the least representative samples pooled, with raked probability weights included*.

SI-22 Supplementary Information References

Franco, Annie, Neil Malhotra, Gabor Simonovits, and L. J. Zigerell. "Developing standards for post-hoc weighting in population-based survey experiments." *Journal of Experimental Political Science* 4, no. 2 (2017): 161-172.

Kolenikov, Stanislav. "Updates to the ipfraking ecosystem." *The Stata Journal* 19.1 (2019): 143-184.

Miratrix, Luke W., Jasjeet S. Sekhon, Alexander G. Theodoridis, and Luis F. Campos. “Worth Weighting? How to Think About and Use Weights in Survey Experiments.” *Political Analysis* 26, no. 3 (2018): 275–91. https://doi.org/10.1017/pan.2018.1.

Winship, Christopher, and Larry Radbill. "Sampling weights and regression analysis." *Sociological Methods & Research* 23.2 (1994): 230-257.