METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Assessing the Efficacy of a Participant-Vetting Procedure to Improve Data-Quality on Amazon's Mechanical Turk

Emilio D. Rivera[1], Benjamin M. Wilkowski[1], Aaron J. Moss[2], Cheskie Rosenzweig[2,3], Leib Litman[2,4]

[1] *Department of Psychology, University of Wyoming, Laramie, WY, USA.* **[2]** *Prime Research Solutions, Queens, NY, USA.* **[3]** *Department of Clinical Psychology, Columbia University, New York, NY, USA.* **[4]** *Department of Psychology, Lander College, Flushing, NY, USA.*

Methodology, 2022, Vol. 18(2), 126–143, https://doi.org/10.5964/meth.8331

Received: 2022-02-09 • **Accepted:** 2022-06-08 • **Published (VoR):** 2022-06-30

Handling Editor: Marcelino Cuesta, University of Oviedo, Gijón, Spain

Corresponding Author: Emilio D. Rivera, 1000 East University Avenue, Department of Psychology, Department 3415, University of Wyoming, Laramie, WY, 82071, USA. E-mail: erivera4@uwyo.edu

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

In recent years, Amazon's Mechanical Turk (MTurk) has become a pivotal source for participant recruitment in many social-science fields. In the last several years, however, concerns about data quality have arisen. In response, CloudResearch developed an intensive pre-screening procedure to vet the full participant pool available on MTurk and exclude those providing low-quality data. To assess its efficacy, we compared three MTurk samples that completed identical measures: Sample 1 was collected prior to the pre-screening's implementation. Sample 2 was collected shortly following its implementation, and Sample 3 was collected nearly a full-year after its implementation. Results indicated that the reliability and validity of scales improved with the implementation of this prescreening procedure, and that this was especially apparent with more recent versions. Thus, this prescreening procedure appears to be a valuable tool to help ensure the collection of high-quality data on MTurk.

Keywords

online data collection, screening solutions, vetting procedures, data quality

Since launching in 2005, Amazon's Mechanical Turk (MTurk) has become an increasingly important component to research in many social sciences. MTurk is an online platform

composed of between 100,000 and 200,000 active people worldwide (Robinson et al., 2019). People on MTurk (i.e., Turkers) complete "Human Intelligence Tasks" (HITs; e.g., surveys or other computerized tasks) in exchange for immediate monetary compensation; Robinson et al., 2019). Since launching, the use of MTurk samples has been increasingly popular in many social science fields (Kennedy et al., 2020).

A major reason that MTurk has become a staple in modern research is due to the speed of data collection and access to a broader population compared to traditional undergraduate samples. While Turkers are somewhat younger, more liberal, and tend to be better educated than the general population as a whole, they are considerably more representative than college samples (Paolacci & Chandler, 2014). The ability to sample niche groups also aids in research that is difficult or impossible with more traditional recruitment techniques.

Despite MTurk's advantages, there are several limitations. The current article is focused on four main issues that have been heavily discussed in past research: inattention, trolling, automated bots, and foreign workers (Kennedy et al., 2020). Each of these issues has displayed worrying trends over the last two years, with researchers increasingly finding that data collected through MTurk displays the tell-tale signs of being compromised. Recently, however, CloudResearch has instituted a more rigorous pre-screening procedure.

The current research aims to test whether data quality concerns may be mitigated by this vetting procedures. To this end, we compare data from an open sample of Turkers to data collected from CloudResearch's "Approved Participants" list. We then examine how data screening solutions that measure in-survey performance might be used to further remove any problematic participants at the study-level.

Data Quality Concerns

Inattention

It has been suggested that Turkers often give inadequate attention to study materials (e.g., Kennedy et al., 2020). Because they are not directly monitored, Turkers may divide their attention between the HIT and other stimuli. This can lead to random or repetitive response patterns that compromise the validity and reliability of the data (Meade & Craig, 2012). While this concern has drawn considerable attention, research has indicated that Turkers may actually be more attentive than online college samples (Hauser & Schwarz, 2016). Thus, this issue is not new and applies to traditional research populations as well.

Trolling

Another issue is the possibility of "trolling"– i.e., intentionally providing insincere responses to be provocative or humorous (Lopez & Hillygus, 2018). Some researchers suggest that trolling in MTurk samples has increased recently, as endorsement of low

base rate responses is implausibly high and increasing (Ahler et al., 2021). However, others suggest that it is difficult to distinguish trolling from other problems, such as inattention (Moss et al., 2021). Regardless, like inattentive responding, trolling would compromise both the reliability and validity of data.

"Bots" and Automated Programs

In 2018, concerns were raised that automated programs (i.e., "bots") were being used to complete large amounts of surveys (Kennedy et al., 2020). Researchers remain concerned about this possibility, given that bots could not only compromise their data's validity and reliability but also result in financial losses. Fortunately, recent research suggests that this concern may have been misplaced, as problematic responses continue even when reCAPTCHA technology is used to filter out automated programs, and many problematic Turkers show evidence of uniquely human processing (Moss et al., 2021). Instead, evidence suggests that the original "bot-panic" may have been due to foreign workers.

Foreign Workers

A major concern for MTurk researchers is their (limited) ability to restrict participation to a target location, such as the United States. While MTurk provides screening tools to limit IP addresses to specific locations, these restrictions can be circumvented (Kennedy et al., 2020). This is particularly problematic for researchers who are restricting studies to American samples, where pay rates are substantially higher than in other regions (Kennedy et al., 2020). This can compromise the data's reliability and validity in multiple fashions. For example, some foreign workers may only have a limited understanding of English (Moss et al., 2021). Further, many psychological phenomena of course vary across cultures, and unknowingly sampling individuals from outside the target culture would introduce systematic error variance.

Consequences for Data Quality

Each of these issues can potentially compromise the reliability and validity of data. For example, inattention or poor English comprehension can result in repetitive response patterns. When scales include reverse-scored items, these items can become *positively* correlated with negative-scored items, compromising reliability.

An even more serious concern is that these issues can compromise *validity* – such that scales no longer measure the target construct. A scale's construct validity is typically assessed by examining the surrounding *nomological network* (Cronbach & Meehl, 1955). If the above issues are severe, the established nomological network surrounding a measure can disappear. Consistent with this, Chmielewski and Kucker (2020) observed a significant loss in the reliability and validity of unscreened data, compared to data that had been screened for inattention.

Data Screening Solutions

Several data-screening methods have been developed to partially address the above issues, such as attention-check items (Oppenheimer et al., 2009). For example, researchers can directly instruct participants to select a particular response and discard participants who fail to provide it.

Another frequent recommendation is to include at least one open-ended question (Kennedy et al., 2020). Bots and participants who are not English-proficient will often provide non-sense answers to such questions. Inattentive participants will often leave such items blank or provide one-word answers, and thus researchers are typically encouraged to use questions that could not reasonably be answered in one word.

A more recent innovation is to include a number of questions with low base-rate response options (e.g., *"Have you ever been convicted of a federal crime"*; Lopez & Hillygus, 2018). The selection of any one low base-rate response is not necessarily indicative of the discussed issues, as it is of course possible that an occasional participant is telling the truth. Nonetheless, the selection of multiple low-base rate responses provides more definitive evidence of such problems.

Finally, reCAPTCHA software can be used to root out bots (von Ahn et al., 2008). With a mouse click, reCAPTCHA differentiates the movement patterns produced by human respondents from those produced by automated programs. However, this solution only screens out a small fraction of bad data (Moss et al., 2021).

While each of these solutions helps reduce common issues, they are not fool-proof. First, they are time-consuming to implement. If researchers recruit their target sample size and then discard 20% of the data, the resulting study is underpowered. They can of course replace these participants through additional data-collection, but this takes additional time. Past research also demonstrates that any single screening item is unlikely to remain effective for long, as Turkers learn what kinds of measures to look out for and show improvements in performance over time (Hauser & Schwarz, 2016).

CloudResearch's Pre-Screening Solution

In 2020, CloudResearch's MTurk Toolkit was developed to help researchers sample from participants on MTurk with added flexibility and demographic targeting capabilities (Litman et al., 2020). Specifically, CloudResearch has been vetting worker accounts on MTurk since spring 2020, implementing a platform-level solution to data quality problems. The goal of this vetting process is to identify Turkers who are likely to provide quality responses to surveys and separate them from Turkers who show signs of the issues reviewed above. As of this writing, CloudResearch has vetted more than 95,000 worker accounts (~90% of active accounts, Litman et al., 2020) and placed more than 60,000 Turkers onto an Approved List. Each month about 5,000 new worker accounts are vetted.

Method

In Summer 2020, two of the current authors collected Sample 1 on MTurk without using the screening procedures described above. As we report below, it was apparent that the resulting data was highly problematic. After correspondences with CloudResearch, we decided to re-run the study using CloudResearch's Approved Participants, to compare the quality of the data collected across samples. After additional improvements to the Pre-Screening procedure, we re-ran the study again roughly one year later to examine possible further improvements. Importantly, all studies used previously-validated measures of political ideology, whose psychometric properties and nomological networks have been established. As such, a comparison of the samples helps to illustrate the efficacy of CloudResearch's screening procedure. We also included several in-survey data-screening solutions in Samples 2 and 3 to examine if data from participants who pass these checks are even more reliable and valid than the data resulting from CloudResearch's pre-screening procedure alone.

Open Science Practices

We report all measures relevant to current concerns, how we determined our sample size, and data exclusions. Verbatim method files, data, and analytic code can be accessed at the Supplementary Materials section. A few additional measures were included that have not been as thoroughly validated in prior research. We disclose all such measures in full in the verbatim method files, but do not focus on them here. In all studies, we aimed to collect a final analyzable sample of 250 participants (Schönbrodt & Perugini, 2013). In all studies, we over-sampled in order to meet this goal after data exclusions.

Participants

In each sample, participants were recruited through MTurk. To be eligible, participants needed at least a 95% HIT approval rate. In Samples 2 and 3, all participants had to additionally pass CloudResearch's vetting procedure (described below).

Sample 1

This sample was collected in June of 2020 and included 274 (192-males, 82-females; M_{age} = 37.11, *SD* = 10.56) participants. Our sample identified primarily as being either White/non-Hispanic (46.35%) or Black/African-American (45.26%) and conservative-leaning (*M* = 5.18, *SD* = 1.75). Please note that this does not match the typical demographics of Turkers (Paolacci & Chandler, 2014), raising concerns about the accuracy of demographic reports. Due to not being within the target population of the original study (i.e., current resident of United States), 17 participants were removed from the sample and not included in analyses.

Sample 2

This sample was collected shortly after the implementation of the prescreening procedure in July of 2020, and it included 306 (162-females, 143-males, 1-unidentified; M_{age} = 39.57, *SD* = 13.76) participants. Our samples identified primarily as being White/non-Hispanic (70.59%) and liberal-leaning (*M* = 3.69, *SD* = 1.87). Note that these demographics more closely resemble the typical demographics of MTurk samples (Berinsky et al., 2012). Due to not being within the target American population, 7 participants were removed from analyses.

Sample 3

This sample was collected in August of 2021, after improvements to the vetting procedure. It included 300 (170-females, 129-males, 1-unidentified; M_{age} = 42.27, *SD* = 13.14) participants. Participants primarily identified as White/non-Hispanic (72.67%) and liberal-leaning (*M* = 3.44, *SD* = 1.81). No participants were removed due to residing outside of the United States.

CloudResearch Vetting Process

Part of CloudResearch's MTurk vetting process is an attention instrument (Litman et al., 2020), which checks for participant attention and basic English comprehension. Participants are presented with a target word, four response-option words, and asked which response-option is most similar to the target. The stimuli were created using an associative semantic network model, which assigns weights to word-pairs based on a large corpora of English language texts. The resulting questions are ones where the correct response option is highly associated with the target word (e.g., *apologize-mistake*) and the distractor response options have low associations with the target word (e.g., *apologize-particle*). Because items are randomly selected from a large library, participants are less likely to learn the correct responses, share them online, or create scripts that can respond automatically.

CloudResearch's vetting process also includes items that check for mischievous responding, items that check for bots (e.g., reCAPTCHA), and technical checks that look for evidence of other automated solutions (Litman et al., 2020). In tandem, CloudResearch also monitors the number of times workers are added by MTurk Toolkit users to a "universal exclude" list and uses this data to further identify questionable respondents. This vetting process appears to effectively identify some of the most problematic participants on MTurk (Moss et al., 2021). In this study, we examine how well the Approved List improves data quality.

Procedures and Measures

Participants completed the following scales in a random order, with item order within each instrument also randomized. A full list of items and other related material can be found in the Supplementary Materials section.

Social Dominance Orientation (SDO)

Participants in both samples completed the 16-item SDO-7 scale (Ho et al., 2015). This instrument contains an equal number of positively-scored items and negatively-scored items. Participants rated their level of favorability towards each item on a 7-point Likert-scale (1–*Strongly oppose* to 7–*Strongly favor*). It has been found to be internally consistent in past research (α = .89; Ho et al., 2015).

Right Wing Authoritarianism (RWA)

Participants also completed the 22-item RWA scale (Altemeyer, 2006). Items includes 12 negatively-scored items. Participants were instructed to rate their agreement with each item on a 9-point Likert-scale (-4–*Very strongly disagree* to 4–*Very strongly agree*). This scale has been found to be internally consistent in past research (α = .90; Saunders & Ngo, 2017)

Aggression-Submission-Conventionalism (ASC)

Participants also completed the ASC (Dunwoody & Funke, 2016) measure of RWA. It contains an equal number of positive and negatively-scored items. Participants rated their agreement with each item on a 7-point Likert-scale. It has been found to be internally consistent in past research (α = .85; Dunwoody & Funke, 2016)

Left-Wing Authoritarianism (LWA)

Participants also completed the 22-item LWA scale (Conway et al., 2018). It includes 10 negatively-scored items. Participants rated their agreement with each item on a 7-point Likert-scale. It has been found to be internally consistent in past research (α = .89; Conway et al., 2018)

World Views

Participants also completed the 10-item Dangerous World Beliefs scale (DWB) and the 14-item Competitive Jungle World Beliefs scale (CJWB) (Duckitt et al., 2002). These scales each contain an equal number of positively and negatively-scored items. All scales ask participants to indicate their level of agreement using a 7-point Likert-scale. The DWB scale (α = .80) and the CJWB scale (α = .84) have been found to be internally consistent in past research (Duckitt et al., 2002).

Fundamental Social Motives Inventory (FSMI)

Participants also rated their agreement with 12 items that measured self-protection and disease avoidance motives (Neel et al., 2016). The Self-Protection and Disease-Avoidance subscales contain 1 and 3 negatively-scored items (respectively). Participants rated that agreement with each item on a 7-point Likert-scale. These scales have been found to be internally reliable in past research (αs > .80; Neel et al., 2016).

Measure of Moral Motives (MMM)

Participants completed the 30-item Measure of Moral Motives Scale (Janoff-Bulman & Carnes, 2016), which measures 6 motives. Here, we focus on Social-Order and Social-Justice motives, which have been linked to Conservativism and Liberalism (respectively). Participants rated their agreement with each item on a 7-point Likert-scale. The MMM contains no negatively-scored items. The MMM scales have all been found to be internally reliable in past research (αs > .70; Janoff-Bulman & Carnes, 2016).

Moral Foundation Sacredness Scale (MFS)

Participants completed the 24-item Moral Foundations Sacredness Scale (MFS; Graham & Haidt, 2012). Items ask participants to indicate how much they would need to be paid to engage in actions typically viewed as immoral. It has 5 subscales (Loyalty, Authority, Purity, Harm, & Fairness). The Binding Foundations (Loyalty, Authority, & Purity) are related to Conservativism; while the Individualizing Foundations (Harm & Fairness) are modestly associated with Liberalism. It has no negatively-scored items. Each scale (αs > .65) has been found to be at least modestly reliable (Graham & Haidt, 2012).

Generalized Prejudice Items

To measure generalized dimensions of prejudice, participants responded to warmth thermometer items for 44 different social groups. Twenty-four items were modeled after items from Duckitt and Sibley (2007) to measure derogated groups (e.g., *Immigrants*), dangerous groups (e.g., *Violent Criminals*), and dissident groups (e.g., *Protestors*). We selected items that seemed to represent the same dimensions in a 2020 American sample, and added a few unique items. We also added items to measure prejudice against conservative groups (Brandt et al., 2014). Participants were instructed to rate their warmth towards each group along a 0–100 continuum (0–*No Warmth* to 100–*Extremely Warm*). We reverse-scored all items to form a measure of prejudice.

In-Survey Screening Items

Participants in Samples 2 and 3 were asked to complete four Attention Check items, which directly instructed them to select a particular response (e.g., *I will select six now and show I'm paying attention*). These were distributed across four scales.

Participants in Samples 2 and 3 were also asked two direct validation questions (i.e., "*Were you paying attention when answering questions?*", "*Were you telling the truth?*"). Participants were assured that they would be paid regardless of their answers. Past research indicates that participants who indicate inattention or deception provide lower-quality data (Lopez & Hillygus, 2018).

In Samples 2 and 3, we also asked several open-response items pertaining to their height, state of residency, and annual income to screen for nonsense responses. We also added four low base-rate items (e.g., *Are you in a gang?*). The endorsement of two or more low base-rate responses is indicative of trolling or other problems (Lopez & Hillygus, 2018).

Finally, we asked Samples 2 and 3 participants to complete two reCAPTCHA items, at the beginning and end of the survey. This helped to screen out responses provided by automated software.

Results

Initial Screening of Participants

We first screened Sample 2 and 3's responses to the In-Survey Screening items. We used this to create two subsamples, Sample 2-Additional Screening (S2-AS) and Sample 3-Additional Screening (S3-AS), who displayed in-survey evidence of providing high-quality data.

In Sample 2, 43 of the 306 total participants (14%) were excluded—7 because they were outside the United States; 29 for failing two or more attention checks; 5 for directly indicating they provided low-quality data; and 2 for endorsing two or more low base-rate items. After their removal, S2-AS included 263 participants (Male = 114, Female = 148, Non-Conforming = 1; M_{age} = 40.98, *SD* = 14.05). This sub-sample identified primarily as being White/non-Hispanic (73.76%) and liberal-leaning (*M* = 3.67, *SD* = 1.90).

In Sample 3, 22 of the total 300 participants were excluded (7.3%). This suggests a significant improvement in the Approved Worker Pre-Screening procedure over time, as fewer participants were excluded than Sample 2, χ^2 (1, *N* = 606) = 7.14, *p* = .008. 19 were excluded for failing two or more attention-check items, and 3 for directly indicating their data was low-quality. After their removal, S3-AS included 278 participants (Male = 118, Female = 159, Non-Conforming = 1; M_{age} = 42.89, *SD* = 13.16). They primarily identified as White/non-Hispanic (74.10%) and liberal-leaning (*M* = 3.47, *SD*=1.81).

Correlation Between Positive and Reverse-Scored Items

Perhaps the clearest indication of problematic responses by Sample 1 lies in the correlation between responses to positive and negative-scored items. Seven previously validated scales included negative-scored items, which typically correlate negatively with positive-

ly-scored items on the same instrument. In contrast to this typical finding, the average response to reverse-scored items was *positively* correlated with the average response to positively-scored items in Sample 1 (see Table 1). In Samples 2 and 3, these correlations returned to their typical negative direction (Sample 2 mean r = -.43; Sample 3 mean r = -.56). In the sub-samples that passed additional screening, these inverse correlations were slightly stronger (S2-AS mean r = -.52; S3-AS mean r = -.67).

Table 1

Correlation Between Responses to Positive and Negative-Scored Items

Variable	S1	S2-Full	S2-AS	S3-Full	S3-AS
1. SDO	0.39*	-0.71*	-0.78*	-0.74*	-0.84*
2. RWA	0.77*	-0.44*	-0.54*	-0.59*	-0.68*
3. LWA	0.77*	-0.03	-0.14*	-0.25*	-0.37*
4. ASC	0.65*	-0.37*	-0.47*	-0.45*	-0.59*
5 .DWB	0.58*	-0.44*	-0.55*	-0.65*	-0.78*
6. CJWB	0.50*	-0.45*	-0.52*	-0.57*	-0.67*
7. FSMI: Harm Avoidance	0.30*	-0.43*	-0.50*	-0.61*	-0.71*
8. FSMI: Disease Avoidance	0.35*	-0.57*	-0.62*	-0.60*	-0.71*

Note. S1 = Sample. S2-Full = All Sample 2 participants. S2-AS = Sample 2 participants who passed additional screening.
*$p < .05$.

Inter-Item Reliability

We next examined inter-item reliability. As noted above, negatively-scored items are included on 7 scales. When we properly scored these scales in Sample 1, many of them proved to be highly unreliable with alphas as low as .18 (see Table 2). When we improperly scored these questionnaires by leaving negative-scored items in their "raw" form, their reliability coefficients actually increased. This lends further supports the idea that many Sample 1 participants provided repetitive responses. In Samples 2 and 3, this issue was eliminated. All scales were internally reliable when properly-scored, with αs > .82 (see Table 2). Improperly scoring them lowered each scale's reliability (Sample 2 mean α = .55; Sample 3 mean α = .45). Additional screening helped to make this pattern slightly more apparent, but the effect was relatively minimal (unscreened mean α = .889; additional screening mean α = .894). It is important to note that this issue is not apparent in scales containing no negative-scored items. For these scales, a consistent tendency to favor a certain side of the scale can result in an internally reliable scale. This underscores the utility of including negatively-scored items in online data collection.

Table 2

Inter-Item Reliabilities (Cronbach's Alpha)

	Sample 1		S2 - Full		S2 - AS		S3 - Full		S3 - AS	
Variable	**Proper**	**Raw**	**Proper**	**Raw**	**Proper**	**Raw**	**Proper**	**Raw**	**Proper**	**Raw**
Scales Containing Negative-Scored Items										
1. SDO	0.71	0.90	0.95	0.45	0.95	0.28	0.95	0.35	0.95	0.05
2. RWA	0.41	0.96	0.92	0.71	0.92	0.65	0.93	0.62	0.94	0.51
3. LWA	0.18	0.94	0.83	0.82	0.84	0.77	0.86	0.74	0.87	0.66
4. Harm-Avoidance	0.63	0.80	0.85	0.65	0.87	0.63	0.91	0.66	0.92	0.62
5. Disease-Avoidance	0.35	0.79	0.86	0.01	0.87	0.21	0.87	0.10	0.89	0.64
6. ASC	0.44	0.93	0.88	0.66	0.92	0.61	0.87	0.54	0.88	0.35
7. DWB	0.20	0.87	0.86	0.47	0.91	0.34	0.91	0.17	0.82	0.32
8. CJWB	0.46	0.89	0.89	0.60	0.87	0.51	0.89	0.42	0.89	0.18
Scales Without Negative-Scored Items:										
9. BJW	0.87	--	0.94	--	0.94	--	0.94	--	0.93	--
13. Social Order	0.83	--	0.89	--	0.89	--	0.86	--	0.87	--
14. Social Justice	0.73	--	0.90	--	0.90	--	0.90	--	0.90	--
16. Harm	0.87	--	0.87	--	0.87	--	0.89	--	0.85	--
17. Fairness	0.85	--	0.68	--	0.68	--	0.76	--	0.73	--
18. Ingroup Loyalty	0.86	--	0.76	--	0.76	--	0.80	--	0.78	--
19. Authority	0.85	--	0.72	--	0.74	--	0.82	--	0.80	--
20. Purity	0.86	--	0.74	--	0.75	--	0.73	--	0.69	--

Construct Validity

The most serious problem is when the validity of instruments themselves is compromised, such that they no longer measure the target construct. Classically, construct validity is assessed by examining a "nomological network" that contains positive correlations with measures of similar constructs, and weaker correlations with measures of more unrelated constructs (Cronbach & Meehl, 1955). Fortunately, many aspects of the nomological network for the current measures are well-established. Table 3 summarizes the most important aspects of this network. Broadly, the typical pattern of correlations is not apparent in Sample 1. Instead, the pattern of correlations can be better interpreted as a consistent repetitive response tendency extending across multiple measures. In Samples 2 and 3, however, the more typical network re-appears.

In contrast to previous research (Janoff-Bulman & Carnes, 2016; Sibley & Duckitt, 2008), Conservatism was related to greater Social-Justice motives in Sample 1; while Liberalism was related to greater prejudice against many groups (including *liberal* groups). This pattern is consistent with repetitive response tendencies (i.e., favoring the high side of the scale would result in reports of Conservatism, social justice motives, and

less-prejudice). While other findings are more consistent with past research, they could also reflect a repetitive response tendency.

Table 3

Correlations Between Conservativism and Previously-Associated Constructs in all Samples

	Political Conservativism (vs. Liberalism)				
Variable	**S1**	**S2-Full**	**S2-AS**	**S3-Full**	**S3-AS**
1. MMM: Social Justice	0.15*	-0.50*	-0.56*	-0.45*	-0.52*
2. LWA	-0.14*	-0.46*	-0.50*	-0.57*	-0.59*
3. MFS: Harm	0.15*	-0.12*	-0.15*	-0.06	-0.09
4. MFS: Fairness	0.19*	>-.01	-0.02	-0.02	-0.04
5. RWA	0.27*	0.67*	0.70*	0.57*	0.60*
6. ASC	0.25*	0.61*	0.73*	0.44*	0.46*
7. SDO	0.41*	0.57*	0.63*	0.53*	0.56*
8. DWB	-0.02	0.39*	0.42*	0.36*	0.37*
9. CJWB	0.29*	0.36*	0.41*	0.17*	0.21*
10. MMM: Social Order	0.46*	0.49*	0.52*	0.47*	0.48*
11. MFS: In-Group Loyalty	0.20*	0.22*	0.25*	0.18*	0.18*
12. MFS: Authority	0.20*	0.20*	0.20*	0.16*	0.15*
13. MFS: Purity	0.15*	0.16*	0.18*	0.10	0.10
16. Generalized Prejudice: Illegal Behaviors	-0.32*	0.24*	0.27*	0.32*	0.30*
17. Generalized Prejudice: Derogated	-0.19*	0.18*	0.18*	0.15*	0.16*
18. Generalized Prejudice: Liberal Dissidents	-0.17*	0.60*	0.65*	0.64*	0.67*
19. Generalized Prejudice: Conservative Dissidents	-0.34*	-0.41*	-0.44*	-0.35*	-0.38*
20. Generalized Prejudice: Dangerous	-0.34*	-0.02	0.02	0.17*	0.19*
21. Generalized Prejudice: Foreign	-0.18*	0.34*	0.37*	0.32*	0.34*

Note. See Table 1 notes for abbreviations.

*$p < .05$.

The typical nomological network re-appeared in Samples 2–3. Social-Justice motives were related to Liberalism (Janoff-Bulman & Carnes, 2016); while most forms of prejudice were related to Conservatism (Duckitt & Sibley, 2007). These samples also replicated an important exception to the latter pattern: Liberals actually expressed greater prejudice toward *conservative* groups (Brandt et al., 2014).

Finally, Sample 1 was unable to replicate more nuanced distinctions between RWA and SDO, but Samples 2-3 were able to replicate such distinctions (see Table 3). Compared to RWA, SDO exhibited stronger correlations in Samples 2–3 with being male (Pratto et al., 1997), lower concern with Harm/Fairness (Sinn & Hayes, 2017), greater prejudice against derogated groups (Duckitt & Sibley, 2007), and more Competitive

Worldviews (Duckitt et al., 2002). By contrast, RWA exhibited stronger correlations (than SDO) in Samples 2–3 with Religiosity (Heaven & Connors, 2001), the 'Binding' Foundations (Sinn & Hayes, 2017), and Dangerous Worldviews (though its typical correlation with prejudice against dangerous groups was not replicated; Duckitt et al., 2002). None of these findings were apparent in Study 1.

Additional screening of Samples 2 (S2-AS) and 3 (S3-AS) made all of these patterns slightly more apparent, but the effect was relatively minimal.

Discussion

MTurk has provided researchers with a reliable, efficient, and affordable participant pool that is more representative than college student samples. Recently, however, the quality of data provided by MTurk samples has come under scrutiny. To overcome these issues, CloudResearch implemented a vetting procedure.

To investigate the impact of these vetting procedures, three samples were collected using largely identical measures at different times: before implementation, immediately after implementation, and one year after implementation. Screening items were also added to Samples 2–3. While 12% of Sample 2 and 7% of Sample 3 were removed (e.g., because of failed attention checks), this is significantly lower than rates in unvetted MTurk samples (Kennedy et al., 2020). The reduction in exclusion rates from Samples 2 to 3 suggests that the vetting procedure has become more effective over time.

Data Quality

To examine the efficacy of the vetting procedures, we compared data quality across the three samples. Importantly, the vetted Samples 2–3 exhibited a marked increase in reliability and validity compared to the unvetted Sample 1. Additional screening resulted in only a very modest increase in reliability/validity.

Scale Reliability

Internal reliability indices indicate how strongly items on a scale are related to one another. Sample 1 exhibited problematically low internal reliability on previously-validated scales. Surprisingly, we observed *increased* reliability when negatively-scored items were improperly scored. Sample 1 responses appear to be driven by repetitive response patterns, rather than by veridical reports of the target construct. In the vetted samples, though, this pattern disappeared. All scales were reliable (all αs > .80), and improper scoring decreased reliability. This suggests that the vetting procedure reduced problems such as inattentive or non-English-proficient responses.

Construct Validity

Beyond reliability, another important concern is whether a scale measures the intended construct. To assess this, we examined the established nomological network surrounding our measures. Within Sample 1, the established nomological network was not apparent. Social Justice Motives were related to Conservatism. Liberalism was related to prejudice against many groups (including *liberal* groups!), and more nuanced distinctions between RWA and SDO could not be made. After vetting, however, evidence of scale validity returned in Samples 2–3. Social Justice Motives resumed their typical correlation with Liberalism (Janoff-Bulman & Carnes, 2016), and Conservatism resumed its normal correlation with prejudices against many groups (Sibley & Duckitt, 2008), except for conservative groups (Brandt et al., 2014). More nuanced distinctions between RWA and SDO also reappeared (Duckitt & Sibley, 2007; Duckitt et al., 2002). Thus, the vetting procedure helps to ensure scales successfully measure their intended construct.

Impacts of Additional Screening

The removal of participants through in-survey screening items (e.g., attention checks, open-ended questions) led to an increase in reliability and validity. However, this increase was modest, at best. For example, the removal of problematic participants slightly increased the magnitude of convergent correlations with conservativism in Sample 2 (average before removal: $r = .35$; average after removal: $r = .38$) and Sample 3 (average before removal: $r = .32$; average after removal: $r = .34$). Thus, removal was generally beneficial, but extremely modest compared to the vetting procedure.

In summary, our results provide evidence for the utility of CloudResearch's vetting procedures. Nonetheless, we still recommend that researchers screen out participants who fail to meet pre-specified criteria. In many contexts (e.g., with small effect sizes or low-frequency outcomes), even small numbers of problematic participants can lead to inaccurate conclusions (Litman et al., 2020). Moreover, this prescreening procedure removes only the most problematic and clearly fraudulent actors. Few participants are perfect all the time, and only in-survey screening items can exclude participants who are merely having an off day.

Practical Recommendations

Given the current findings, what practical recommendations can we make? For example, some readers may be tempted to discount all MTurk research conducted during the same time period as Sample 1. However, Sample 1 is better regarded as an example of what *can go wrong* without proper precautions; and we cannot assume that such issues immediately generalize to all MTurk studies during this time. While CloudResearch's vetting procedure is one effective solution, it is not the only one. If research conducted during this period contains evidence of proper precautions (e.g., attention-checks, reCAPTCHA,

evidence of valid/reliable measures), then its data is likely more trustworthy than Sample 1. Without such evidence, however, readers should be attentive to the possibility that such data is indeed compromised.

Another important question is how long this vetting procedure method will remain effective. The process of maintaining data quality on MTurk is an ongoing process (Hauser & Schwarz, 2016). While CloudResearch's solution has demonstrated continued effectively over the first year, this initial success does not necessarily speak to its continued success in the years to come. As such, the effectiveness of this vetting procedure should be periodically re-examined in the coming years, and it may be necessary to develop new procedures if Turkers' behavior changes substantially in response to it.

Finally, future research should also examine the impact of this vetting procedure on other types of errors. Our focus here was solely on its impact on Measurement Error (i.e., reliability & validity). However, the Total Survey Error framework (e.g., Groves & Lyberg, 2010) clearly indicates that this is only one type of error. The impact of this vetting procedure on other types of error (e.g., sampling error) is currently unknown, and future research is needed to address this.

Conclusions

Since its founding, MTurk has provided social scientists with a consistent and relatively inexpensive data source. However, issues such as inattention and foreign workers can compromise data quality. CloudResearch instituted a vetting procedure aimed at identifying participants who provide more reliable data. The current results suggest that this vetting procedure increases reliability and validity of data. While we continue to recommend in-survey screening items, this vetting procedure is one viable means of collecting reliable and valid data.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: Some of the authors of this manuscript are employed at CloudResearch (i.e., Prime Research Solutions). CloudResearch provides online research tools and services, including tools that allow researchers to run studies on Mechanical Turk. CloudResearch's MTurk ToolKit was used to source Mechanical Turk participants, and the CloudResearch database was used to query some of the data. All data was analyzed and reported by researchers not affiliated with CloudResearch.

Data Availability: Both Data and Materials are freely available at Supplementary Materials.

Supplementary Materials

The supplementary materials provided include verbatim method files, data, and analytic code that can be accessed at (see Index of Supplementary Materials below).

Index of Supplementary Materials

Rivera, E. D., Wilkowski, B. M., Moss, A. J., Rosenzweig, C., & Litman, L. (2022). *Supplementary materials to "Assessing the efficacy of a participant-vetting procedure to improve data-quality on Amazon's Mechanical Turk"* [Method files, data, analytic code]. OSF. https://osf.io/5bvfk/?view_only=f0a8ae59bd2349f28eff6e894399d99d

References

Ahler, D. J., Roush, C. E., & Sood, G. (2021). The micro-task market for lemons: Data quality on Amazon's Mechanical Turk. *Political Science Research and Methods*. Advance online publication. https://doi.org/10.1017/psrm.2021.57

Altemeyer, B. (2006). *The authoritarians*. B. Altemeyer. https://theauthoritarians.org/

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. https://doi.org/10.1093/pan/mpr057

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science, 23*(1), 27–34. https://doi.org/10.1177/0963721413510932

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological & Personality Science, 11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Conway, L. G., III, Houck, S. C., Gornick, L. J., & Repke, M. A. (2018). Finding the Loch Ness monster: Left-wing authoritarianism in the United States. *Political Psychology, 39*(5), 1049–1067. https://doi.org/10.1111/pops.12470

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Duckitt, J., & Sibley, C. G. (2007). Right wing authoritarianism, social dominance orientation and the dimensions of generalized prejudice. *European Journal of Personality, 21*(2), 113–130. https://doi.org/10.1002/per.614

Duckitt, J., Wagner, C., Du Plessis, I., & Birum, I. (2002). The psychological bases of ideology and prejudice: Testing a dual process model. *Journal of Personality and Social Psychology, 83*(1), 75–93. https://doi.org/10.1037/0022-3514.83.1.75

Dunwoody, P. T., & Funke, F. (2016). The Aggression-Submission-Conventionalism Scale: Testing a new three factor measure of authoritarianism. *Journal of Social and Political Psychology, 4*(2), 571–600. https://doi.org/10.5964/jspp.v4i2.168

Graham, J., & Haidt, J. (2012). Sacred values and evil adversaries: A moral foundations approach. In P. Shaver & M. Mikulincer (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 11–31). APA Books. https://doi.org/10.1037/13091-001https://doi.org/10.1037/13091-001

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Heaven, P. C. L., & Connors, J. R. (2001). A note on the value correlates of social dominance orientation and right-wing authoritarianism. *Personality and Individual Differences, 31*(6), 925–930. https://doi.org/10.3758/s13428-015-0578-z

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R., Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO_7 scale. *Journal of Personality and Social Psychology, 109*(6), 1003–1028. https://doi.org/10.1037/pspi0000033

Janoff-Bulman, R., & Carnes, N. C. (2016). Social justice and social order: Binding moralities across the political spectrum. *PLoS One, 11*(3), Article e0152479. https://doi.org/10.1371/journal.pone.0152479

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods, 8*(4), 614–629. https://doi.org/10.1017/psrm.2020.6

Litman, L., Rosen, Z., Ronsezweig, C., Weinberger-Litman, S. L., Moss, A. J., & Robinson, J. (2020). *Did people really drink bleach to prevent COVID-19? A tale of problematic respondents and a guide for measuring rare events in survey data.* medRxiv. https://doi.org/10.1101/2020.12.11.20246694

Lopez, J., & Hillygus, D. S. (2018). Why so serious? Survey trolls and misinformation. *SSRN.* https://doi.org/10.2139/ssrn.3131087

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085

Moss, A. J., Rosenzweig, C., Jaffe, S. N., Gautam, R., Robinson, J., & Litman, L. (2021). *Bots or inattentive humans? Identifying sources of low-quality data in online platforms.* PsyArXiv. https://doi.org/10.31234/osf.io/wr8ds

Neel, R., Kenrick, D. T., White, A. E., & Neuberg, S. L. (2016). Individual differences in fundamental social motives. *Journal of Personality and Social Psychology, 110*(6), 887–907. https://doi.org/10.1037/pspp0000068

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184–188. https://doi.org/10.1177/0963721414531598

Pratto, F., Stallworth, L. M., & Sidanius, J. (1997). The gender gap: Differences in political attitudes and social dominance orientation. *British Journal of Social Psychology, 36*(1), 49–68. https://doi.org/10.1177/0963721414531598

Robinson, J., Rosenzweig, C., Moss, A. J., & Litman, L. (2019). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PLoS One, 14*(12), Article e0226394. https://doi.org/10.1371/journal.pone.0226394

Saunders, B. A., & Ngo, J. (2017). The right-wing authoritarianism scale. *Encyclopedia of Personality and Individual Differences, 1*(4), 1–4. https://doi.org/10.1016/j.jesp.2009.03.009

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Sibley, C. G., & Duckitt, J. (2008). Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review, 12*(3), 248–279. https://doi.org/10.1177/1088868308319226

Sinn, J. S., & Hayes, M. W. (2017). Replacing the moral foundations: An evolutionary-coalitional theory of liberal-conservative differences. *Political Psychology, 38*(6), 1043–1064. https://doi.org/10.1111/pops.12361

von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCaptcha: Human-based character recognition via web security measures. *Science, 321*(5895), 1465–1468. https://doi.org/10.1126/science.1160379