Running Head: REPUTATION ON AMAZON MECHANICAL TURK

Reputation as a sufficient condition for data quality on Amazon Mechanical Turk

Eyal Peer [a], Joachim Vosgerau [b], Alessandro Acquisti [c]

[a] Graduate School of Business Administration, Bar-Ilan University, Ramat-Gan, Israel, 52900; eyal.peer@biu.ac.il (corresponding author).

[b] School of Economics and Management, Tilburg University; j.vosgerau@uvt.nl

[c] Heinz College, Carnegie Mellon University; acquisti@andrew.cmu.edu

Published Version:

Behav Res (2014) 46:1023–1031
DOI 10.3758/s13428-013-0434-y

Reputation as a sufficient condition for data quality on Amazon Mechanical Turk

Abstract

Data quality is one of the major concerns of using crowdsourcing web sites such as Amazon Mechanical Turk (MTurk) to recruit participants for online behavioral studies. We compared two methods for ensuring data quality on MTurk: attention check questions (ACQs) and restricting participation to MTurk workers with high reputation (above 95% approval ratings). In Experiment 1, we found that high reputation workers rarely failed ACQs and provided higher quality data than low reputation workers; ACQs improved data quality only for low reputation workers, and only in some of the cases. Experiment 2 corroborated these findings and also suggested that more productive high reputation workers produce the highest quality data. We conclude that sampling high reputation workers can ensure high quality data without having to resort to using ACQs ,which may lead to selection bias if participants who fail ACQs are excluded post-hoc.

Key words: Online research; Amazon Mechanical Turk; Data quality; Reputation

Reputation as a sufficient condition for data quality on Amazon Mechanical Turk

An increasing number of social scientists are capitalizing on the growth of crowd-sourced participant pools such as Amazon Mechanical Turk (MTurk). One of the main issues that has been occupying researchers using this pool of participants is data quality (e.g., Goodman, Cryder, & Cheema, 2012). Recent studies have shown that various forms of attention check questions (ACQs) to screen out inattentive respondents or to increase the attention of respondents are effective in increasing the quality of data collected on MTurk (e.g., Aust, Diedenhofen, Ullrich & Musch, 2012; Buhrmester, Kwang , & Gosling, 2011; Downs, Holbrook, Sheng, & Cranor, 2010; Oppenheimer, Meyvis, & Davidenko, 2009). Such ACQs usually include "trick" questions (e.g., "have you ever had a fatal heart attack?", Paolacci, Chandler, & Ipeirotis, 2010) or instructions which ask respondents to answer a question in a very specific way (e.g., skip it or enter prescribed responses). The main objective of these ACQs is to filter out respondents who are not paying close attention to the experiment's instructions. Additionally, including such ACQs in an experiment can help to increase or ensure participants' attention, as they do not know when to expect another trick question as the experiment progresses (Oppenheimer et al., 2009).

The use of ACQs can be particularly effective when researchers have no prior knowledge about participants' motivation and capacity to read, understand, and comply with research instructions. MTurk, however, offers researchers information about participants' past performance, or reputation, in form of approval ratings. Every time a participant (a.k.a., "worker") on MTurk completes a task (a.k.a., "Human Intelligence Task", or "HIT"), the provider (a.k.a., "requester") of that task can approve or reject a worker's submission. Rejecting

a worker's submission also involves denying that worker her or his payment for completing the HIT and reflects badly on that worker's account. Furthermore, it can reduce the variety of HITs that a worker can work on in the future, because requesters can demand workers to have a minimum number of previously approved HITs to be eligible for their HIT. While MTurk does not disclose individual workers' approval ratings to requesters, it allows requesters to set a minimum qualification for workers to view and complete a HIT (e.g., that 95% of their previous HITs were approved). The main objective of setting this kind of qualification is to try to ensure that the responses collected for the study would be reliable and credible, and would enable the research to reach its objectives.

In this paper, we compare the effectiveness of these two methods for ensuring data quality on MTurk: restricting samples to MTurk workers with high reputation (e.g., 95% or more of previous HITs approved) versus using ACQs to screen out inattentive workers and/or to increase their attention. We compare both methods in terms of validity, reliability, and replicability of research findings.

Ensuring data quality: Attention checks vs. approval ratings

Having participants pass ACQs or sampling those who have a high reputation could both improve data quality but may also bear unintended consequences. Restricting participation to MTurk workers with high reputation reduces the size of the population from which a sample is drawn, thereby potentially prolonging the time needed to reach a required sample size. Furthermore, sampling bias may result if workers with high reputation differ from those with low approval ratings on dimensions other than attention and willingness to comply with experimental instructions.

Using ACQs to screen out inattentive respondents, on the other hand, diminishes sample size and can lead to unequal experimental cell sizes and selection bias if responses are excluded after data collection is completed (Oppenheimer et al., 2009). Furthermore, ACQs might backfire. For example, ACQs such as "Have you ever, while watching TV, had a fatal heart attack?"—to which an attentive respondent must respond with 'never' (Paolacci et al., 2010)—may cause reactance on the respondents' part. An attentive respondent might take offense by the surveyor's implicit assumption that s/he does not pay enough attention, and react by being less thorough in subsequent responding or by providing outright wrong answers. While other ACQs can be less offensive (e.g., researchers can explain, in the ACQ, why it is important for them to make sure participants are reading the instructions), adding an unrelated question (such as ACQs) can potentially disrupt the natural flow of a study. If ACQs are necessary to obtain high quality data, than a relatively small disruption in the study's flow is probably negligible. However, if ACQs do not improve data quality (or do so only for certain groups of MTurk workers – such as those with low reputation), than the use of ACQs should probably be discouraged to avoid potential reactance and selection bias.

To compare the effectiveness of both methods, we ran two experiments on MTurk in which we orthogonally varied MTurk worker's reputation (below vs. above 95%) and the use of ACQs in the study (mandatory vs. absent). We assessed data quality in terms of reliability, validity, and replicability. For reliability, we asked participants to fill out several validated scales measuring individual differences (on personality, self-esteem, need for cognition, and social desirability). We used the social desirability scale also to assess data quality in terms of validity – assuming that more socially desirable responses are less valid. Finally, following Paolacci et al. (2010), we assessed data quality in terms of replicability of well-known effects.

In the first experiment, we focused on comparing high vs. low reputation workers and manipulated the use of ACQs to assess the contribution of each method to increasing data quality. In the second experiment, we replicated the first experiment's results, using different (and less familiar) ACQs, and also examined differences between workers with different productivity levels (i.e., those who completed less vs. more previous HITs on MTurk).

Experiment 1

Method

Sampling and participants. During 10 days, we sampled U.S. respondents from two populations on MTurk: workers with above 95% approval ratings (high reputation), and workers with below 95% approval ratings (low reputation). The cutoff of 95% was chosen because—as the default setting in MTurk—it is used by many researchers. The cutoff, however, is arbitrary, and higher or lower cutoffs can be used for distinguishing high versus low reputation workers. The responses of 694 workers, 458 with a high reputation and 236 with a low reputation, were obtained. A power analysis shows that with these sample sizes effect sizes of d = .25 and above will be detected in about 90% of the cases. To verify workers' reputation, we asked them to report their approval ratings. While 91.1% of the high reputation workers confirmed to have a higher than 95% approval rating, 36.0% of the low reputation workers claimed to have an approval rating of above 95%, χ^2 (5) = 263.3, p < .001. Rather than doubting the validity of MTurk's qualification system, we believe that these participants - intentionally or not - misreported their approval ratings. No statistically significant differences in either gender (χ2 (3) = 2.04, p = .56) or age (F (3, 690) = 1.59, p = .19) were found across groups (see Table 1).

Design. About 70% of each sample (high and low reputation workers) were administered ACQs and the remaining were not. ACQ conditions were oversampled because we wanted to

compare responses of those who failed to those who passed ACQs (see samples' sizes in Table 1).

Table 1: Demographics by group in Experiment 1.

	High reputation		**Low reputation**	
	With ACQs	No ACQs	With ACQs	No ACQs
N	302	156	177	59
Mean of Age (*SD*)	32.12 (11.27)	33.96 (12.21)	32.40 (11.49)	35.00 (13.99)
% female	48.0%	43.6%	51.4%	47.5%

Procedure. Participants were invited to complete a survey about personality. The survey started with demographic questions, followed by the Ten-Item Personality Inventory (TIPI, Gosling, Rentfrow, & Swann, 2003), Rosenberg's 10-items Self-Esteem Scale (RSES, Rosenberg, 1979), the short 18-items form of the Need for Cognition scale (NFC, Cacioppo, Petty, & Kao, 1984), and the short 10-item form of the Social Desirability Scale (SDS, Fischer & Fick, 1993). All measures used five-point Likert scales with end points *strongly disagree* (1) and *strongly agree* (5), except for the SDS, which used a binary scale with *agree* (1) and *disagree* (0). Participants were then asked to complete a classic anchoring task (Tversky & Kahneman, 1974): They first entered the last two digits of their phone number, then indicated if they thought the number of countries in Africa was larger or smaller than that number, and finally estimated the number of countries in Africa.

In the ACQ condition, three ACQs were included in different parts of the survey. The Instructional Manipulation Check (IMC, Oppenheimer et al., 2009) was inserted right after the demographic questions. Participants were asked "Which sports do you like?", but hidden in a lengthy text were instructions to ignore the question and simply click on next. The second ACQ (after the NFC questionnaire, before the anchoring task) asked—among other unobtrusive questions—"While watching TV, have you ever had a fatal heart attack?" (Paolacci et al., 2010). The last ACQ at the end of the survey asked participants "What was this survey about?", preceded by instructions to not mark 'Personality' but instead choose 'Other' and type 'Psychology' in the text box (adapted from Downs et al., 2011). Participants were paid 50 cents.

Results

Attention Check Questions (ACQs). We compared the rates of failing ACQs between high and low reputation workers. As can be seen in Table 2, only 2.6% of high reputation workers failed at least one ACQ compared to 33.9% of low reputation workers (χ^2 (1) = 89.46, $p < .001$). For example, 0.4% of high reputation workers indicated that they had a fatal heart attack while watching TV, while 16.4% of low reputation workers claimed to have suffered such a deadly incident. Given that almost all high reputation workers (97.4%) passed all ACQs, for the subsequent analyses we created five comparison groups: high reputation workers who either received (and passed) ACQs or did not receive ACQs, and low reputation workers who either passed all ACQs, failed ACQs at least once, or did not receive any ACQs. The sample sizes for these groups are given in Table 3.

Table 2: Proportion of participants who failed ACQs.

Reputation	**Passed** (all ACQs)	**Failed** (at least one ACQ)	**# of ACQs failed**		
			1	**2**	**3**
High	294 (97.4%)	8 (2.6%)	8 (2.6%)	0	0
Low	117 (66.1%)	60 (33.9%)	35 (19.8%)	14 (7.9%)	11 (6.2%)

Table 3: Data quality measures among high and low reputation workers.

		High reputation		**Low reputation**		
		Passed ACQs	No ACQs	Passed ACQs	Failed ACQs	No ACQs
N		302	156	117	60	59
Cronbach alpha	SDS	$.629_{a}$	$.698_{a}$	$.471_{b}$	$.242_{b}$	$.557_{ab}$
	RSES	$.936_{a}$	$.934_{ad}$	$.912_{ad}$	$.825_{bc}$	$.889_{cd}$
	NFC	$.952_{a}$	$.947_{ad}$	$.891_{ad}$	$.759_{bc}$	$.863_{cd}$
SDS mean percent (SD)		44.87 (21.5)	45.71 (23.7)	48.63 (18.9)	53.0 (17.3)	49.83 (21.2)
Anchoring effect size (r)		$.198_{a}$*	$.183_{a}$ *	$.280_{a}$*	$-.046_{b}$	$.049_{b}$
Average percent of midpoint marked on scale items (SD)		19.28 (14.1)	20.78 (14.06)	25.12 (17.04)	34.21 (26.56)	27.61 (21.08)

Statistics with different subscripts differ at $p < .05$ (Cronbach alpha pairwise comparisons tested following Hakstian & Whalen (1976) with Bonferroni-correction for post-hoc testing)
* Significantly different from 0 at $p < .05$

Reliability. We regarded internal consistency (Cronbach's alpha) of established scales as evidence of data quality and compared it between the different groups of workers. First, we compared the reliability scores for the SDS, RSES and NFC scales of high vs. low reputation workers (we did not examine the reliability of the TIPI scales because each scale only had two items). High reputation workers produced higher reliability scores on all three scales (.635, .935, .950, for SDS, RSES, and NFC, respectively) compared to low reputation workers (.452, .887, .865, respectively). Using Hakstian & Whalen's (1976) test for statistical significance between independent reliability coefficients, we found that the differences in reliabilities between high and low reputation workers were statistically significant for all three scales, χ^2 (1) = 13.75, 19.86, 62.60, $p < .001$. Participants who had failed ACQs produced lower reliability scores on all three scales (.563, .821, .761, respectively) compared to those who had passed (.601, .931, .942) or not received ACQs (.666, .923, .932), χ^2 (1) > 13.75, $p < .001$. When testing all possible pairwise comparisons among the five groups, using Bonferroni's correction for post-hoc comparisons[1], lower reliabilities were found among low reputation workers who either failed or not received ACQs compared to high reputation workers (whether they had or had not received ACQs, see Table 3).

Social Desirability. We regarded socially desirable responding as evidence of lower data quality. Comparing the five groups, we found statistically significant differences, F (4, 689) = 2.52, $p = .04$, $\eta^2 = .014$. Low reputation workers who had failed ACQs had the highest SDS scores, while high reputation workers showed the lowest scores ($p < .05$; see table 3). However, none of the pairwise comparisons was statistically significant (after a Bonferroni's correction, see Table 3).

[1] That is, multiplying the p-values by the number of possible comparisons, which were, in this case and in all other analyses reported for this study, 10.

Anchoring task. Following Paolacci et al. (2010) and Oppenheimer et al. (2009), we regarded replicability of well-established effects as evidence for high quality data. Numerous studies have shown that answering a hypothetical question about a clearly arbitrary anchor (e.g., the last two digits of one's phone number) influences subsequent unrelated number estimates (e.g., Tversky & Kahneman, 1974). We expected high reputation workers to be more likely to show the classic anchoring effect than low reputation workers because inattentive respondents are more likely to be distracted during the task, which should weaken an anchoring effect. The last two digits of phone numbers and the number of African countries showed the expected positive correlation—evidence of an anchoring effect—among high reputation workers (with and without ACQs) and among low reputation workers who had passed the ACQs, but not among low reputation workers who did not receive ACQs or had failed them (see Table 3). Bonferroni-corrected post-hoc comparisons showed that the differences between these correlations were statistically significant ($p < .05$).

Central tendency bias. To test whether workers differed in their tendency to mark the midpoint of scales regardless of the questions asked, we computed for each participant the relative frequency with which they had marked "3" on the five point scales in the TIPI, RSES and NFC. An ANOVA on this central tendency bias ratio showed significant differences between the groups, F (4, 689) = 12.76, $p < .001$, $\eta^2 = .07$. As can be seen in Table 3, there was no difference in central tendency bias between high reputation workers who did or did not receive ACQs ($p = 1.0$). Among low reputation workers, those who had passed ACQs showed a significantly greater central tendency bias than those who had failed ACQs ($p = .006$). The

difference between low reputation workers who had passed ACQs and those who did not receive ACQs was not statistically significant (p = .31, all p-values are Bonferroni-corrected).

Discussion

The results of Experiment 1 suggest that workers' reputation can predict data quality: high reputation workers were found to provide higher quality data compared to low reputation workers. High reputation workers rarely failed ACQs (97.4% passed them), and their responses resulted in higher reliability scores for established measures and showed lower rates of socially desirable responding. High reputation workers also exhibited the classic anchoring effect whereas low reputation workers did not. Low reputation workers, in contrast, were found to be more likely to cross off the midpoint of scales regardless of the question asked (central tendency bias).

ACQs did improve data quality, but for low reputation workers only, and only in some of the cases. For the RSES and NFC scales, reliability scores among low reputation workers who had passed ACQs were just as high as scores obtained from high reputation workers who had either passed or not received ACQs. For the SDS scale, however, even low reputation workers who had passed all ACQs produced a significantly lower reliability on that measure. Similarly, ACQs helped improve data quality among low reputation workers in terms of replicability. Low reputation workers who had passed ACQs showed the classic anchoring effect (as did high reputation workers regardless of having received or not received ACQs), whereas low reputation workers who had either failed or not received ACQs failed to produce the expected effect. Finally, low reputation workers showed higher levels of central tendency bias, independently of whether they had received, passed, or failed ACQs.

More importantly though, ACQs did not seem to have any effect whatsoever on the data quality of high reputation workers. The responses of high reputation workers produced high scale reliabilities whether ACQs were used or not, showed the same (low) degree of socially desirable responding, exhibited almost identical effect sizes in the anchoring task, and displayed the same (relatively low) level of central tendency bias. This lack of differences in all of the measures we used strongly suggests that ACQs (or, at least, the ACQs used in this study) do not have an effect on high reputation workers. Such a null effect, however, can only be meaningfully interpreted when the study is adequately powered to detect small effects (Greenwald, 1975). Experiment 1 with almost 700 participants in total would have detected effects of $d = .25$ and above with a probability of 90%. It is hence unlikely that differences among high reputation workers who did or did not receive ACQs actually existed but were not observed in Experiment 1. Rather, the results suggest that ACQs are generally ineffective in improving data quality among high reputation workers who produce very high quality data to begin with.

However, although the ACQs we used in this experiment did not improve data quality, other ACQs may do so. The fact that almost all high reputation workers passed the ACQs suggests that it may be that high reputation workers are familiar with these specific (and common) ACQs (the ACQs that we used in Experiment 1 have been available to researchers for several years now, e.g., the IMC was published in 2009). If familiarity is the cause for the high passing rate of ACQs, novel and unfamiliar ACQs may increase data quality for high reputation workers the same way as they do for low reputation workers. Experiment 2 was designed to test that possibility.

Experiment 2

Experiment 2 employed the same design and measures as Experiment 1 with two exceptions: First, we replaced the ACQs used in Experiment 1 with novel ACQs that we designed ourselves (after soliciting examples from colleagues). Second, in addition to workers' reputation, we orthogonally manipulated workers' productivity. Worker productivity refers to the number of HITs that a worker has previously completed on MTurk. Similar to worker reputation (percent of approved HITs), MTurk allows researchers to specify how many HITs workers must have previously completed to view and complete their HIT. There seems to be high variance in workers' productivity levels. For example, about half of the participants in Experiment 1 indicated that they had completed more than 250 HITs, and about 10% said they had completed more than 5,000 HITs. A worker's productivity—just like a worker's reputation—may be a predictor for data quality, such that highly productive workers may be more likely to produce high quality data than less productive workers. That could be the case because a) highly productive workers are workers who are more intrinsically motivated to complete HITs to the satisfaction of requester, b) highly productive workers represent 'good' workers that stayed on MTurk while 'bad' workers dropped out over time, and c) highly productive workers are more experienced in answering survey questions and thus produce higher quality data.

Experiment 2 served three purposes. First, to see whether the findings of Experiment 1 would replicate, second, to test whether novel and unfamiliar ACQs would improve data quality for high reputation workers, and third, to test whether worker productivity would have the same effect on data quality as worker reputation.

Method

Sampling. During 10 days, we sampled MTurk workers (who did not take part in Experiment 1) from the U.S. with either high or low reputation (above 95% vs. less than 90%

previously approved HITs), and with either high or low productivity levels (more than 500 HITS vs. less than 100 HITs completed). Different from Experiment 1, in Experiment 2 we manipulated both factors reputation and productivity in such a way that there was a gap in between manipulated levels. This way, we avoided MTurk workers with similar reputation/productivity levels (e.g., 95.1% vs. 94.9% approved HITs or 501 vs. 499 completed HITs) being categorized into different groups. As a consequence, it should be easier to detect actual differences in data quality as a function of worker reputation/productivity. The cutoffs for productivity were chosen based on the distribution of self-reported productivity levels in Experiment 1 (about 25% indicated they had completed less than 100, and about 30% said they had completed more than 500 HITs).

Sampling was discontinued when an experimental cell had reached about 250 responses or after 10 days. While we were able to collect responses from 537 high reputation workers in less than two days, we only obtained responses from 19 low reputation workers in 10 days. After two days of very slow data collection, we tried to increase the response rate by re-posting the HIT every 24 hours (so it would be highly visible to these workers) and increasing the offered payment (from 70 to 100 cents for a 10 minutes survey). Unfortunately, both attempts were unsuccessful. We thus decided to focus only on high reputation workers and on the impact of productivity levels and ACQs on data quality. The obtained sample size allows for detecting effect sizes of at least $d = .25$ with a power of about 80%.

Participants. We collected responses from a total of 537 MTurk workers with high reputation (95% or above), 268 with low productivity (100 or less previous HITs) and 269 with high productivity (500 or more previous HITs). Both groups of high vs. low productivity included similar ratios of males (61.5% vs. 58.6%), χ^2 (1) = .43, $p = .51$, but workers from the

high productivity groups were somewhat older than those in the low productivity group (M_{high} = 34.36, SD_{high} = 12.45 vs. M_{low} = 32.08, SD_{low} = 12.62), t (499) = 2.04, p = .04, d = .18. As expected, high productivity workers reported having completed a much larger number of HITs than low productivity workers (M_{high} = 10,954.78, SD_{high} = 38,990.67 vs. M_{low} = 138.64, SD_{low} = 151.49), t (499) = 4.38, p < .001, d = .39. Interestingly, many workers from the low productivity group (about 43%) claimed to have completed more than 100 HITs – a fact that should have prohibited them from taking our HIT. Some (about 24%) of these claimed to have completed more than 250 HITs – an over-report that can be hardly ascribed to simple oversight or memory error. Lastly, high productivity workers reported having, on average, a slightly higher ratio of previously approved HITs than low productivity workers ((M_{high} = 99.40, SD_{high} = .76 vs. M_{low} = 99.21, SD_{low} = 1.14), t (434) = 2.1, p = .04, d = .20.

*Design***.** Participants in each group were randomly assigned to either receive (novel) ACQs or not. As in Experiment 1, we oversampled the condition that included ACQs (in a ratio of about 67:33).

Procedure. As in Experiment 1, MTurk workers were invited to complete a survey about personality for 70 cents. Participants first completed the TIPI, followed by the 10-item version of the SDS, the 10-item version of the RSES, and the 18-item version of the NFC scale. In the last page of the survey, participants were asked to indicate their gender and age, and to estimate approximately how many HITs they had completed in the past and how many of those were approved (in contrast to Experiment 1, these questions did not include pre-defined options but used an open text-box in which participants entered their responses, allowing us to get more granular data).

Participants in the ACQ conditions were asked to answer three additional questions (three novel ACQs): the first one presented participants with a picture of an office in which six people were seated, and asked them to indicate how many people they see in the picture. Hidden within a lengthy introduction were instructions to workers to not enter "6" but instead enter "7" to show that they had indeed read the instructions. Any response other than 7 was coded as failing this ACQ. The second new ACQ was embedded in the middle of the NFC scale in the form of a statement that read: "I am not reading the questions of this survey". Any response other than "strongly disagree" was coded as failing this ACQ. The last novel ACQ consisted of two questions that asked participants to state whether they "would prefer to live in a warm city rather than a cold city" and whether they "would prefer to live in a city with many parks, even if the cost of living was higher." Both questions were answered on 7-point Likert scales with end points *strongly disagree* (1) and *strongly agree* (7). Participants were instructed, however, not to answer the question according to their actual preferences but to mark "2" on the first question and then add 3 to that value and use the result (i.e., 5) as an answer to the second question. Any deviating responses were coded as failing this ACQ.

Results

Attention-Check Questions. As can be seen in Table 4, among those who received the ACQs (about 2/3 of each group), 80.3% of high productivity workers passed all of them compared to 70.9% of low productivity workers, χ^2 (3) = 12.63, p = .006. As in Experiment 1, we classified workers of each productivity group according to whether they had passed all ACQs, had failed at least one of the ACQs, or did not receive ACQs at all (see Table 5 for groups' sizes).

Table 4. Rates of passing/failing unfamiliar ACQs in Experiment 2.

Productivity	**Passed** (all ACQs)	**Failed** (at least one ACQ)	**# of ACQs failed**		
			1	**2**	**3**
High	150 (83.3%)	30 (16.7%)	26 (14.4%)	3 (1.7%)	1 (.6%)
Low	127 (70.9%)	52 (29.0%)	33 (18.4%)	16 (8.9%)	3 (1.7%)

Reliability. As in Experiment 1, we regarded high internal reliability as evidence for high data quality. However, we could not (as we did in Experiment 1) compare reliabilities between high and low reputation workers because we were unable to sample enough low reputation workers. As an alternative, we decided to compare the reliability of the measures used in this study (SDS, RSES and NFC) to their conventional coefficients as reported in the literature: Fischer and Fick (1994) report a reliability of .86 for the short form of SDS with a sample of 309 students; Cacioppo et al. (1984) report a reliability of .90 that was obtained from 527 students; and Robins, Hendin and Trzesniewski (2001) report a reliability of .88 for the RSES among 508 students. We compared the reliability obtained from our MTurk groups to these scores using the Hakistan & Walen (1976) test for significance of differences between independent reliability coefficients. In all analyses, we employed the Bonfferoni's correction method and multiplied p-values by the number of possible comparisons. We found that all groups showed a significantly lower reliability for the SDS compared to the reliability reported in the literature, χ^2 (1) > 15.6, p < .01. However, reliabilities for the RSES and NFC scales were not significantly lower than those reported in the literature (ps > .05). In fact, for some of the cases reliabilities were higher than those reported in the literature, especially among high productivity workers and those who passed ACQs (see Table 5).

Comparing high and low productivity workers, we found that high productivity workers produced higher reliability scores for the SDS, RSES and NFC scales (.70, .931, .951 vs. .576, .910, .912, respectively). These differences were statistically significant for all three scales, χ^2 (1) = 7.15, 4.23, 21.27; p = .0075, .039, p < .001, respectively, suggesting that high productivity workers produced higher quality data. When comparing the three groups who had passed, failed or not received ACQs, we found no statistically significant differences in the reliability scores of the SDS, but we did find statistically significant differences in the RSES and NFC scales, χ^2 (2) = 3.38, 18.84, 7.61; p = .18, p < .001, p = .022, respectively. In the two scales that showed statistical differences (RSES and NFC), participants who had passed ACQs showed higher reliability scores compared to those who had failed or not received ACQs (.938 vs. .897 and .888 for the RSES and .946 vs. .917 and .927 for the NFC scale). However, the scores between those who had failed versus those who did not receive ACQs were not statistically different for either the RSES or the NFC scale, χ^2 (1) = .18, .42; p = .67, 51, respectively.

Table 5. Internal reliability and social desirability scores for the groups in Study 2.

	Low productivity			**High productivity**		
ACQ	**Passed**	**Failed**	**None**	**Passed**	**Failed**	**None**
N	127	52	89	150	30	89
SDS	.586	.441	.608	.741	.72	.605
RSES	.930	.888	.881	.945	.907	.888
NFC	.929	.872	.898	.954	.962	.941
SDS mean percent (SD)	47.64 (20.49)	54.81 (18.94)	45.96 (20.60)	47.53 (24.98)	47.00 (25.75)	46.29 (21.45)

We then examined whether the effect of adding (novel) ACQs occurred both within high and low productivity workers. We compared the reliability scores of the three ACQ groups within each productivity group (which are given in Table 5). Among the low productivity groups, we found no statistical difference for the SDS, but we did find significant differences for the RSES and the NFC scale, χ^2 (2) = 1.86, 7.66, 6.74; p = .39, .02, .03 respectively. Among the high productivity groups, we did not find statistical differences for the SDS or the NFC, but we did find significant differences for the RSES, χ^2 (2) = 4.27, 2.56, 12.62; p = .12, .28, .002, respectively. This suggests that the aforementioned overall effect of ACQs was mostly driven by differences among low productivity workers.

Social desirability. As in Experiment 1, we regarded lower levels of socially desirable responses as a proxy for higher data quality. We calculated for each participant the percent of socially desirable responses according to the SDS (the averages of the SDS percent are reported in Table 5 for the productivity and ACQ groups). An ANOVA on the SDS mean percent scores with productivity and ACQ conditions showed no statistically significant effect for productivity, ACQ, or their interaction, F (1, 2, 2, respectively, 531) = 1.3, 1.24, 1.03, p = .25, .29, .36, η^2 = .002, .005, .004, respectively.

Central tendency bias. To measure participants' tendency to mark the midpoint of the scale, we computed for each participant the relative frequency with which they had marked "3" on the five point scales in the TIPI, RSES and NFC. An ANOVA on this *central tendency bias* score showed a significant effect for ACQs, F (2, 531) = 6.04, p =.003, η^2 = .022, and no significant effects for the level of productivity, F (1, 531) = 3.38, p = .066, η^2 = .006, or the interaction between the two, F (2, 531) = 1.93, p = .15, η^2 = .007. Post-hoc comparisons, using

Bonferroni's correction, showed that those who had passed ACQs were less likely to mark the midpoint of the scales compared to those who had failed the ACQs (M = 1.81 vs. 0.27, SD = .13, .18; p = .009, d = .31). Respondents who did not receive ACQs showed an average score (M = 0.20, SD = .13) that was not significantly different from the other two groups' scores (p > .05).

Discussion

We found corroborating evidence that high reputation workers (whether having previously completed many or few HITs) can produce high quality data. In contrast to Experiment 1 which used familiar ACQs (which may have been ineffective for experienced MTurk workers), Experiment 2 employed three novel ACQs. Even using these novel ACQs did not improve data quality among high reputation workers, replicating the finding from Experiment 1. Together, the findings suggest that sampling high reputation workers appears to be a sufficient condition for obtaining high quality data on MTurk. Note that—as in Experiment 1—this conclusion relies on interpreting a null effect as meaningful, which is possible when samples are adequately powered (Greenwald, 1975). Indeed, our sample had a statistical power of more than 80% to detect differences of at least d = .25. The fact that no differences were found suggests that high reputation workers produce high quality data, irrespective of ACQs.

Additionally, we also found that workers who were more productive (having completed more than 500 HITs, and sometimes much more than that) were less prone to fail ACQs and, in some respects, produced slightly higher data quality than less experienced workers who had completed less than 100 HITs. Moreover, ACQs increased data quality to some extent among low productivity workers but not among high productivity workers. This suggests that sampling highly productive high reputation workers may be the best way to ensure high quality data

without the need of resorting to ACQs. However, one must consider possible drawbacks of including highly productive workers, such as that they might not be totally naïve to the experimental procedure or the questions of the study (see Chandler, Mueller, & Paolacci, 2013, for a discussion of non-naivety amongst MTurk respondents).

General discussion

Data quality is of utmost importance for researchers conducting surveys and experiments using online participant pools such as MTurk. Identifying reliable methods, which ensure and increase the quality of data obtained from such resources is thus important and beneficial. In two studies, we found that one way to ensure high quality data is to restrict sampling of participants to MTurk workers who have accumulated high ratings from previous researchers (or other MTurk requesters). When sampling such high reputation workers, data quality – as measured by scales' reliability, socially desirable responses, central tendency bias, and replicability of known effects – was satisfactorily high. In contrast, low reputation workers seem to pay much less attention to instructions as indicated by a higher failure rate of ACQs, and thus produced data of lower reliability, exhibited more response biases, and showed smaller effect sizes for well-known effects. Our recommendation is to restrict sampling to high reputation (and possibly highly productive) MTurk workers only. In our studies, we used the arbitrary cutoff of 95% to differentiate between workers with high or low reputation levels. Researchers may of course use a stricter cutoff given that the distribution of workers is highly skewed in favor of high reputation workers.

While the first experiment was, in its nature, exploratory, our findings were corroborated in our second experiment, which also helped overcome Experiment 1's main limitation – workers' familiarity with the used ACQs. We found that even when novel and unfamiliar ACQs

were used, high reputation workers showed a high likelihood of passing them (indicating that they do read instructions). In fact, one of the most important findings of our research lies in the null effect that ACQs seem to have on high reputation workers. Whether or not ACQs were used, these high reputation workers provided high quality data, across all of the measures we employed in our studies. Whatever effect ACQs had on MTurk workers was limited to low reputation workers (Experiment 1) or to workers who were less productive (Experiment 2). Even then, the effect was limited to only some of the cases and some of the measures of data quality. Thus, we conclude that sampling high reputation workers is not only a necessary but also a sufficient condition for obtaining high quality data. Using ACQs does not seem to help researchers to obtain higher quality data, despite previous emphasis on this approach (e.g., Aust, Diedenhofen, Ullrich & Musch, 2012; Buhrmester, Kwang , & Gosling, 2011; Downs, Holbrook, Sheng, & Cranor, 2010; Oppenheimer, Meyvis, & Davidenko, 2009). Perhaps ACQs were essential a few years ago, but they do not seem to be essential currently.

Sampling high reputation workers to ensure high data quality without using ACQs provides two advantages. First, when ACQs are used and responses are excluded after data collection, experimental cell sizes may differ and selection bias may occur. Second, ACQs may cause reactance and hamper the natural flow of a study. We did not find evidence for the second advantage, however, it should be noticed that we did not include any measures that were specifically geared towards measuring reactance or survey flow (such as attitudes toward the survey or the researchers).

For our recommendation of not using ACQs but instead restricting sampling to high reputation workers to be beneficial, two things must hold. First, it is important that sampling only high reputation workers would not result in sampling bias, which would be the case if high

reputation workers differed from low reputation workers on dimensions other than paying attention to instructions. In our experiments, we did not find evidence for this to be the case, as high and low reputation workers showed the same distributions of age and gender. It should be noted, however, that we could not assess potential differences in personality traits, self-esteem, and need for cognition scores between high and low reputation workers, because the lower reliability scores and higher levels of central tendency bias among low reputation workers made it impossible to compare these scores to those of high reputation workers. Second, it is important that restricting sampling to high reputation workers does not interfere with response rates. In our experiments, we found no evidence for this to be the case. In fact, the sample in Experiment 1 obtained from low reputation workers after 10 days of data collection was about half the size of the sample obtained from high reputation workers. In Experiment 2, which was conducted a few months later, we were unable to sample a sufficient amount of low reputation workers for our study. Therefore, it seems that restricting samples to high reputation workers does not significantly reduce the pool from which workers are sampled, and will only minimally affect the time needed to reach a specified sample size. In the current state of the MTurk population, sampling only high reputation workers appears to be an effective and efficient method to ensure high data quality on MTurk.

Our studies also point to a possible phenomenon that may be occurring on MTurk, namely that the number (or ratio) of low reputation workers is low and possibly decreasing. In Experiment 1, we found it harder (more time consuming) to sample low than high reputation workers. In Experiment 2, in which we used an even lower cutoff for low reputation workers, it was not possible to collect a sufficient number of responses from this sub-population in the study time frame. Two things may be happening here: MTurk's HITs approval system 'weeds out' bad

workers (i.e., those who perform poorly and do not satisfy requesters' needs). If true, the entire population of MTurk workers will increasingly consist of only highly reputed and productive workers, which would make MTurk an even more attractive pool for researchers. However, another and less fortunate process might be in play. It is possible that requesters are approving HITs more than they should, thereby increasingly inflating workers' reputation levels. As a consequence, reputation levels would become less indicative of high-quality workers, and ACQs would be needed again to differentiate 'good' from 'bad' workers. Although our studies do not provide conclusive evidence for one or the other, our findings do suggest that the first, and more fortunate, process is more probable. Because high reputation workers generated high quality data, and low reputation workers did not, reputation levels appear to be a reliable indicator of data quality. Further research is needed to investigate whether reputation still predicts data quality in the future or on other crowd sourcing resources for data collection.

Acknowledgments

This research was partially supported by a grant from the NSF (number 1012763), awarded to Alessandro Acqusiti.

References

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2012). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, doi: 10.3758/s13428-012-0265-2.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6*(1), 3–5.

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48(3), 306-307.

Chandler, J., Mueller, P., & Paolacci, G. (2013). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 1-19.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? Screening mechanical turk workers. *Proceedings of the 28th International Conference on Human Factors in Computing Systems* (pp. 2399-2402). ACM.

Fischer, D. G., & Fick, C. (1993). Measuring social desirability: Short forms of the Marlowe-Crowne Social Desirability Scale, *Educational and Psychological Measurement*, 53(2), 417-424.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2012). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making,* doi: 10.1002/bdm.1753

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality, 37, 504-528.*

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1-20. doi: 10.1037/h0076157

Hakstian, R., A., & Whalen, T., A. (1976), A k-sample significance test for independent alpha coefficients, *Psychometrika*, *41*(2), 219-231.

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.

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

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*(2), 151-161.

Rosenberg, M. (1979). *Rosenberg self-esteem scale*. New York: Basic Books.

Tversky, A., & Kahneman D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *211*, 453–458.