Behav Res (2018) 50:1563–1567
https://doi.org/10.3758/s13428-017-0984-5

Assessing and overcoming participant dishonesty in online data collection

Chris Hydock[1]

Published online: 22 December 2017

Abstract Crowdsourcing services, such as MTurk, have opened a large pool of participants to researchers. Unfortunately, it can be difficult to confidently acquire a sample that matches a given demographic, psychographic, or behavioral dimension. This problem exists because little information is known about individual participants and because some participants are motivated to misrepresent their identity with the goal of financial reward. Despite the fact that online workers do not typically display a greater than average level of dishonesty, when researchers overtly request that only a certain population take part in an online study, a nontrivial portion misrepresent their identity. In this study, a proposed system is tested that researchers can use to quickly, fairly, and easily screen participants on any dimension. In contrast to an overt request, the reported system results in significantly fewer (near zero) instances of participant misrepresentation. Tests for misrepresentations were conducted by using a large database of past participant records (~45,000 unique workers). This research presents and tests an important tool for the increasingly prevalent practice of online data collection.

Keywords Sampling · Qualification · MTurk · Online participants · Participant honesty

Over the last decade there has been an incredible increase in online data collection through crowdsourcing platforms such as Amazon Mechanical Turk (MTurk; Chandler, Mueller, & Paolacci, 2014). The use of crowdsourcing platforms has provided many benefits to behavioral researchers in psychology, sociology, economics, politics, and business (Goodman et al., 2013; Horton, Rand, & Zeckhauser, 2011; Mason & Suri, 2012). Crowdsourcing platforms provide a sample that is more representative than student participant pools (Berinsky, Huber, & Lenz, 2012), bring large samples to researchers at small schools (Kraut et al., 2004), and are more efficient than other sources of subjects (Berinsky et al., 2012). Despite the great benefits that crowdsourcing platforms offer, many were not originally designed with behavioral research in mind, a fact that is reflected in their default tools and user interfaces (Litman, Robinson, & Abberbock, 2017). The present article offers an accessible solution to one limitation, the inability to easily recruit a sample with specific characteristics. Some solutions exist for those looking to do research with a specific population—for example, females, full-time employees or clinical populations. However, many have shortcomings, in terms of their cost, fairness to workers (Gleibs, 2017), or efficacy. Accordingly, here an alternative system is presented for qualifying participants on the basis of demographic, behavioral, or psychographic dimensions, a system that is flexible, easy to implement, fair, and reliable.

Crowd sourcing and existing qualification systems

Crowd sourcing platforms, which connect those that need work accomplished with those that seek to complete it for compensation, have been a great tool for behavioral researchers. Researchers consistently find results from MTurk replicate data from participant pools (Buhrmester, Kwang, & Gosling, 2011; Litman et al., 2017; Shapiro, Chandler, & Mueller, 2013; Sprouse, 2011). On MTurk, there are those who need tasks (i.e., human intelligence tasks [HITs])

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-017-0984-5) contains supplementary material, which is available to authorized users.

✉ Chris Hydock
ch937@geogetown.edu

[1] Georgetown University, Washington, DC, USA

completed (requesters), and those who complete tasks (workers). By default, workers self-select HITs; because MTurk is relatively representative of the general population (Hitlin, 2016) and many research paradigms rely on random assignment to manipulations, this practice suffices.

However, some methodologies require specific populations. Despite the observations that the community exhibits norms of honesty and accuracy because worker history impacts future ability to work (Rand, 2012; Shapiro et al., 2013; Suri, Goldstein, & Mason, 2011), the fact that workers exhibit similar levels of dishonestly as laboratory participants (Suri et al., 2011), and the fact that 99% of demographic responses were consistent over time (Mason & Suri, 2012; Rand, 2012), there is reason for skepticism. Specifically, the anonymous, online nature of crowdsourcing platforms, coupled with their incentive structures demands careful consideration of population recruitment. In terms of incentives, workers report making minimum wage or less (Hitlin, 2016; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010), leading many to suspect and document worker's misrepresenting their identity (Chandler & Paolacci, 2017; Horton, 2011; Sharpe Wessling, Huber, & Netzer, 2017).

Several solutions exist to address misrepresentation. Experimenters can recruit the full population and drop ineligible participants, but this is costly, especially for niche populations. Experimenters can include eligibility questions before a survey and disqualify ineligible workers without payment. Given the low compensation on these platforms, this disqualification process is ethically questionable (Gleibs, 2017); in fact, a pilot study of 314 workers found that over 99% believed it was unfair. Furthermore, if a HIT is perceived unfair, workers may share information about it on outlets such as turkopticon.com and turkernation.com (Brawley & Pury, 2016; Mason & Suri, 2012).

Another potential solution, the MTurk qualification system (Chandler et al., 2014) enables requesters to grant qualifications that allow only select workers to participate. To use the qualification system, requesters must rely on information they have previously acquired about workers or they must obtain this information by having the worker complete an unpaid qualification test (described as auto-granting). The former approach requires information to be compiled then uploaded to MTurk and limits the possible sample to past participants, which can be problematic given the high turnover rate of workers (Stewart et al., 2015). The latter approach requires participants to essentially complete a small unpaid HIT before they complete the primary HIT, a process that is less than fair to workers (Gleibs, 2017). With either approach, the qualification system also requires use of a clunky API or user interface.

None of this implies that the previously discussed methods are fatally flawed; rather, it is meant to suggest that any method of qualifying workers has pros and cons. Accordingly, the present article describes and tests the preferred method in my lab for the recruitment of specific populations, as a means to suggest that individual researchers adopt the method that best matches their own preferences.

Proposed system

To easily avoid participant dishonesty while remaining fair to participants, the present study proposes and tests a two-step qualification system. Participants were invited to participate in a brief study (e.g., 10–12 demographic questions, taking care to ensure that a single question could not be identified as the qualifier on the basis of the survey topic) for a small amount of compensation. All participants are paid; ineligible participants are excused, eligible participants are forwarded to the "primary" study, for which they will be additionally compensated (via a bonus if using MTurk). Providing bonuses to large numbers of participants has been made increasingly easy via the TurkPrime interface (Litman et al., 2017), but is also possible using the MTurk API (Mueller & Chandler, 2012) or the MTurkR platform (Leeper, 2016).

To test the proposed two-step qualification system, two experiments were conducted that manipulate the qualification method. In the overt qualification condition, the study title, description, and instructions specified that only participants who matched the stated demographic should participate. In the two- step qualification system condition, the method described in the previous paragraph was used. To assess the proportion of participants that misrepresented themselves, experimental results were compared to a database of ~45,000 unique MTurk workers that have participated in previous studies through a single MTurk account (see the supplemental materials for demographic reliability data and demographic descriptors). Analysis of demographic responses from participants with multiple records reveals a high degree of reliability between participants' initial response and later responses; approximately 1% of age or gender responses did not match participants' initial response.

Experiment 1

In Experiment 1, two HITs were posted on MTurk. One HIT was posted using an overt qualification request, that is, it specified only females should participate. The other HIT was posted with the qualification system proposed in this article. The proportion of participants that misrepresented their identity was measured in each condition.

Method

Experiment 1 was run in two stages on MTurk. At time one, an overt qualification was used. A HIT was posted on MTurk with the following heading "**WOMEN ONLY** Short survey on consumer topics," description "Short survey 7 min WOMEN ONLY," and pay for the HIT was listed as $0.50. After accepting the HIT, participants saw "Study on Consumer Decisions. WOMEN ONLY. You risk REJECTION if you past answers contradict this information" and a link they could follow to take a study. Participants then followed a link to an unrelated survey, completed a demographics section, and were then instructed to 'submit' the HIT in MTurk. In the demographics section participants were asked to provide their gender and age; this information was compared to the information contained in our database.

At time two, a qualification test was used. A HIT was posted with the following title "Qualification survey, short survey on consumer topics," description "<1 minute qualification test, eligible participants take full survey for 50 cents," and pay "$0.10." Upon accepting the study, a link was provided to the qualification test. Participants were instructed that they were being paid $0.10 for their responses, regardless of their eligibility for the primary study. Participants then responded to 12 questions in randomized order. The questions were: What is your age? What is your annual income? What is your gender? (Male/female) Which political party do you identify with? What is your race? What is your religion? What is the highest level of education you have completed? What state do you live in? What is your marital status? Which best describes your employment status? Did you vote in the last political election? Do you have children under 18? Responses to the first two questions were numerical text entry, response choices were provided for the remaining questions. Participants that did not respond female were told they were ineligible, but that they should submit the HIT. Participants that responded female were passed through to the same primary survey that was used in the overt qualification condition. In both conditions participants' unique MTurk ID was forwarded to the survey software, allowing comparison of response to the existing database (see TurkPrime for one method: Litman et al., 2017). All HITs used in this article were posted from the same account, following the same procedures. All studies posted from the account use similar titles so as to limit selection biases. Participants that completed the survey in the overt condition were excluded from completing the survey in the qualification condition.

Results and discussion

For the overt qualification condition 298 participants completed the experiment. The mean reported age was 29.7. Of the 298 participants, 289 listed their gender as female, and nine reported their gender as male in the demographics section at the end of the survey. Interestingly, of the nine participants that reported they were male in the demographic section, all were verified as male by the database, suggesting inattention to the warnings, or disregard for the overt qualification request. Next, the database was searched for all 298 participants; 240 of the participants that completed the study were in the database. The database responses indicated that 31 of these 240 (12.9%) had previously reported their gender as male.

For the qualification system condition, 301 participants completed the full experiment; 582 completed the qualification test and 281 were determined to be ineligible. The mean reported age as 32.1. Of the 301 participants, 300 reported their gender to be female in the demographics section, one reported their gender as male. Notably, this respondent necessarily indicated their gender to be female in the qualification test, and the database showed they had previously reported their gender as female. Next the database was searched for the 301 participants, 208 of them were in the database. The database responses indicated that four of these 208 (1.9%) had previously reported their gender as male.

Two chi-square tests were run to determine whether the proportions of participants misrepresenting their identity differed by conditions. The first compared whether the 31 of 240 participants who misrepresented their identity in the overt qualification condition was a statistically greater proportion than the four of 208 who misrepresented their identity in the qualification test condition (according to the database). The percentage of participants who misrepresented their identity was different by condition, $\chi^2(1, N = 448) = 18.69$, $p < .001$. The second, more conservative test compared whether the 22 of 240 who misrepresented their identity (and did not self-report their real gender) in the overt qualification condition was statistically greater than the five of 208 who misrepresented their identity in the qualification test condition (according to either the database or the final demographic questions). The percentage of participants who misrepresented their identity was still different by condition, $\chi^2(1, N = 448) = 9.00$, $p < .01$.

In Experiment 1, the results showed that participants do misrepresent their identity in online platforms, but also that implementing the proposed system can significantly reduce the proportion of misrepresentations. Although it is possible that males are more dishonest than females in this regard (only males had the motive to be dishonest in the overt qualification condition in the experiment), the 12.9% who misrepresented themselves was a nontrivial portion, which suggests that formal qualification tests were necessary.

Experiment 2

Experiment 2 again looked at the proportion of participants who misrepresented their identity when an overt qualification

was used, as compared to when our qualification test is used. In the experiment, the demographic of age (over 35) was used as the qualifying dimension rather than gender.

Method

The methods of Experiment 2 matched those of Experiment 1 except where specified here. In Experiment 2, the first stage of data collection employed the qualification test, and the second employed the overt request. For those completing the qualification test, participants that did not respond over 35 were told they were ineligible. For the overt Qualification test the Title and description read "**AGE OVER 35 ONLY** Short survey on consumer topics," description "Short survey 7 min OVER 35 ONLY."

Results and discussion

For the overt qualification condition 200 participants completed the study, 54% were female. Of the 200 participants, 190 listed their age as over 35, and ten reported their age as less than 35. Of the ten participants that reported they were under 35 in the demographic section, eight participants reported ages matched their database value, suggesting inattention to the warnings, or disregard for the overt qualification request. Two participants were not found in the database. Next, the database was searched for the 200 participants, 124 of the participants that completed the study were in the database. The database responses indicated that 13 of these 124 (10.5%) had previously reported their age as something less than 35.

For the qualification system condition, 193 participants completed the full experiment; 700 completed the qualification test and 507 were told they were ineligible. Of the participants 55.5% were female. Next the database was searched for the 193 eligible participants. Of the 193 participants, 138 had records in the database. Their database responses were compared to the qualification test data and their response in the demographics section of the survey. All 138 responses indicated their age was above 35. Additionally, both the qualification test and demographics section responses corresponded to the database values for all 138 participants.

Two chi-square tests were run to determine whether the proportion of participants misrepresenting their identity differed by condition. The first test compared whether the 13 of 124 participants who misrepresented their identity in the overt qualification condition was a statistically greater proportion than the 0 of 138 who misrepresented their identity in the qualification test condition (according to the database). The percentage of participants who misrepresented their identity was different by condition, $\chi^2(1, N = 262) = 15.22, p < .001$. In the second, a more conservative test compared whether the five of 124 participants who misrepresented their identity (and did not self-report their real age) in the overt qualification condition was statistically greater than the 0 of 138 who misrepresented their identity in the qualification test condition. The percentage of participants who misrepresented their identity was again different by condition, $\chi^2(1, N = 262) = 5.67, p < .05$. These results from Experiment 2 replicated those from Experiment 1. When given the chance to misrepresent their identity, a nontrivial portion of online lab participants did so. However, the proposed qualification system significantly reduced the frequency of such misrepresentation.

General discussion

Over two studies, a small but nontrivial portion of online participants misrepresented their identity for the chance of financial gain. A simple, easy, and fair system, however, was used to significantly reduce the rate of participant misrepresentation. This system serves as a viable option for qualifying participants. It is optimal for scenarios in which no preexisting qualification data are available, or for researchers seeking to avoid the native MTurk qualification system due to technical or ethical reasons.

This research makes several contributions to existing work in online data collection and MTurk in particular. First, it verified past research that showed that participants are willing to misrepresent their identity in order to become eligible for a study, presumably to achieve financial gain. Second, a system to qualify participants was proposed that significantly reduced such misrepresentation. This test was done using a large pool of past participants, data from this pool replicates findings from past studies regarding the composition of MTurk workers (Hitlin, 2016) and the reliability of demographic responses (Mason & Suri, 2012; Rand, 2012). This work also corroborates the recommendations of two articles that offer best practices for MTurk users and also document participant dishonesty (Chandler & Paolacci, 2017; Sharpe Wessling et al., 2017).

A central premise of this proposed system is the fact that participants are paid for answering the questions that qualify them for a study. This contrasts with the standard procedure in many research initiatives, which use unpaid screening questions. However, past research has pointed out the special ethical concerns that are warranted for MTurk, given its status as a convenience sample of respondents who complete studies for incredibly low compensation (Gleibs, 2017). Specifically, the proposed system is based on the belief that, because MTurk workers provide such a great service to the scientific community, so often, and at such low cost (Bohannon, 2011), these participants should be compensated for the questions that are used to determine their eligibility for studies. Also, payment for screening questions will help ensure an ongoing honest relationship.

Finally, for researchers concerned that participants being told they are ineligible may encourage later dishonesty (despite the payment for the eligibility test) could tweak the proposed system by describing the eligibility test as simply a very short survey, and inviting eligible participants for a follow up, rather than telling ineligible participants they were "ineligible."

Conclusion

Altogether, this research helps inform the process of recruiting participants online. It replicates past work on the demographics of MTurk workers and the reliability of their responses. However, it also cautions that proper procedures must be followed to avoid a small but nontrivial portion of participants willing to be dishonest for financial gain. Finally, it proposes and tests an easy, fair, and reliable method of qualifying study participants.

Author note C.H. is now an Assistant Professor of Research at Georgetown University's McDonough School of Business. This research was funded by the university research account of C.H.

References

Berinsky, A., Huber, G., & Lenz, G. (2012). Evaluating online labor markets for experimental research: Amazon.com s Mechanical Turk. *Political Analysis*, *20*, 351–368.

Bohannon, J. (2011). Human subject research: Social science for pennies. *Science*, *334*, 307. https://doi.org/10.1126/science.334.6054.307

Brawley, A., & Pury, C. (2016). Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior*, *54*, 531–546.

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. https://doi.org/10.1177/1745691610393980

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*, 112–130. https://doi.org/10.3758/s13428-013-0365-7

Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*. Advance online publication. https://doi.org/10.1177/1948550617698203

Gleibs, I. H. (2017). Are all "research fields" equal? Rethinking practice for the use of data from crowdsourcing market places. *Behavior Research Methods*, *49*, 1333–1342. https://doi.org/10.3758/s13428-016-0789-y

Goodman, J., Cryder, C., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213–224.

Hitlin, P. (2016). *Research in the crowdsourcing age, a case study.* Washington, DC: Pew Research Center. Retrieved from www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/

Horton, J. (2011). The condition of the Turking class: Are online employers fair and honest? *Economics Letters*, *111*, 10–12.

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

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist*, *59*, 105–117. https://doi.org/10.1037/0003-066x.59.2.105

Leeper, T. J. (2016). *Package "MTurkR": R Client for the MTurk Requester API*. Retrieved from https://github.com/leeper/MTurkR

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433–442. https://doi.org/10.3758/s13428-016-0727-z

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1–23. https://doi.org/10.3758/s13428-011-0124-6

Mueller, P., & Chandler, J. (2012). *Emailing workers using Python (March 3, 2012)*. Retrieved from http://ssrn.com/abstract=2100601

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making*, *5*, 411–419.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*, 213–220. https://doi.org/10.1177/2167702612469015

Sharpe Wessling, K., Huber, J., & Netzer, O. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research*, *44*, 211–230.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*, 155–167. https://doi.org/10.3758/s13428-010-0039-7

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision making*, *10*, 479–483.

Suri, S., Goldstein, D. G., & Mason, W. A. (2011). Honesty in an online labor market. *Human Computation*, *11*, 61–66.