# Are Your Participants Gaming the System? Screening Mechanical Turk Workers

**Julie S. Downs[1], Mandy B. Holbrook[1], Steve Sheng[2], Lorrie Faith Cranor[2,3]**
[1]Social and Decision Sciences, [2]Engineering & Public Policy, and [3]School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{downs, mandy, shengx, lorrie}@cmu.edu

## ABSTRACT

In this paper we discuss a screening process used in conjunction with a survey administered via Amazon.com's Mechanical Turk. We sought an easily implementable method to disqualify those people who participate but don't take the study tasks seriously. By using two previously pilot tested screening questions, we identified 764 of 1,962 people who did not answer conscientiously. Young men seem to be most likely to fail the qualification task. Those that are professionals, students, and non-workers seem to be more likely to take the task seriously than financial workers, hourly workers, and other workers. Men over 30 and women were more likely to answer seriously.

## Author Keywords

screening, Mechanical Turk, survey, crowdsourcing

## ACM Classification Keywords

J.4 [Social and Behavioral Sciences]: Psychology; H.1.2 [User/Machine Systems]: Software psychology.

## General Terms

Experimentation, Human Factors

## INTRODUCTION

Amazon.com's Mechanical Turk (mTurk) is an online marketplace where individuals can perform very small tasks for micro payments, making it an attractive market for researchers to run studies quickly and cheaply through crowdsourcing [1]. Crowdsourcing allows many people to participate with minimal recruitment and administration costs [7]. However, the cash payouts, anonymity and lack of participant accountability may entice people to complete as many tasks as possible without fully engaging in them. If arbitrary clicking pays as well as thoughtful participation, some people may attempt to maximize their profits while minimizing their effort. Indeed, a paper on foreign-language translations found that prolific mTurk users, known as Turkers, performed barely above chance [4]. One

strategy for combating this tendency is to use aggregate or congruent responses from multiple users. However, studies aiming to correlate performance or responses on different tasks rely on meaningful data from each participant. Rather than aggregating noisy data, an alterative strategy would be to develop a reliable method of screening participants to remove the subset of those gaming the system.

## RELATED WORK

Turkers tend to be younger, female, and lower-income than the average Internet user [5]. Payments on mTurk are suggested to follow a reasonable hourly rate, with an example of $8 per hour or about 13¢ per minute [2]. In practice, many mTurk tasks pay much less overall, with the median study paying just 5-10¢ for a task taking "a few minutes," like watching and providing feedback on 3 short (15-second) videos, summarizing a website, and evaluating hypothetical and real market products. Indeed, "wages" this low have been shown to result in lower quality output than could be had for no payment at all, by pure volunteers [3].

Studies using mTurk generate user data quickly and at a low cost, but special consideration needs to go into creating the study materials. As guidance, Kittur and his colleagues provide a set of recommendations for mTurk users to maximize the usefulness of their data [6], some of which will be necessarily limited to certain kinds of tasks. For example, tasks could be designed so that a good-faith effort requires similar or less effort than random responding. However, merely reading the material is a non-negligible amount of work for many surveys. The temptation to choose a convenient response from a multiple-choice set, or to type a superficially appropriate response in a text box, requires considerably less effort than any good-faith response. Thus, other strategies may be effective at identifying respondents not acting in good faith.

Kittur also highlights the need to include items that can be explicitly verified, both to identify appropriate responses and to indicate to the user that responses will be scrutinized [6]. Such items might be mixed in with other items, as in a social desirability scale, to serve as an external indicator that can vouchsafe the trustworthiness of less evaluable responses. Indeed, multiple indicators of suspicious performance are recommended, such as time spent on task, and responses to different types of questions. Here, we

**Table 1. Text of Email and Qualifying Questions**

explore a screening task for an online survey that could not be designed in compliance with mere good-faith incentives.

## RECRUITMENT

An mTurk HIT (Human Intelligence Task) was posted for a 30-minute task [8] with a $4 payment, contingent upon qualification, and 20¢ for those not qualifying. This relatively large payment is roughly equivalent to the federal minimum wage, and corresponds to Amazon's suggested pay rate [2], although it is larger than typical. We chose to pay a more equitable rate both for fair compensation for participants' time and to appeal to a broad array of groups.

## SCREENING TOOL

Some studies identify negligent participants using tricks like, "to show that you are paying attention, please select the third option below," where the suggested option is a clearly incorrect answer. Such transparent tools only catch the most egregious of participants, violate Gricean norms by requiring careful attention to normally predictable information, and set a tone of distrust for the remainder of the task. In contrast, our screening task was designed to appear as a formality, following the logic of the study task. It included 2 questions of varying difficulty, to assess whether there were differential predictors of adhering to strict versus weak criteria. The hallmark of both screening questions was to have a conspicuous distractor. Questions were piloted and refined prior to the current study.

Participants were asked about demographics (age, gender, current occupation) and office work (computer use and participation in conference calls), followed by two qualification questions referring to an email message detailing an upcoming teleconference (see Table 1). The intention was to convey that demographic background or prior participation with teleconferencing may be a prerequisite for qualification, but there was no indication what kind of answers would disqualify. One qualifying question was relatively easy, and could be answered correctly by simply reading the full question and looking up the answer in the email recipient line, with a conspicuous distractor corresponding to the sender's name. The more difficult question required not only interpretation of the question but a close reading of the email text, as the answer was in the body of the email. In this case, the conspicuous distractor corresponded to a piece of information in the signature line of the email. Those looking quickly for answers would likely be inclined to select these distractors.

## RESULTS

The survey design did not require responses in order to continue, and a number of respondents skipped questions. Fifty-five respondents (2.8%) skipped all seven questions,

•     Student (n=482)
•     Professional
    o   Art, Writing and Journalism (n=76)
    o   Education (n=111)
    o   Legal (n=15)
    o   Medical (n=36)
    o   Science, Engineering IT professional (n=296)
    o   Skilled Labor (n=21)
    o   Other Professional (n=187)
•     Financial
    o   Business, Management and Financial (n=172)
•     Hourly
    o   Service (n=69)
    o   Administrative Support (n=123)
•     Not working
    o   Not Currently Working / Unemployed (n=162)
    o   Retired (n=21)
•     Other
    o   Decline to answer (n=22)
    o   Other profession (n=114)
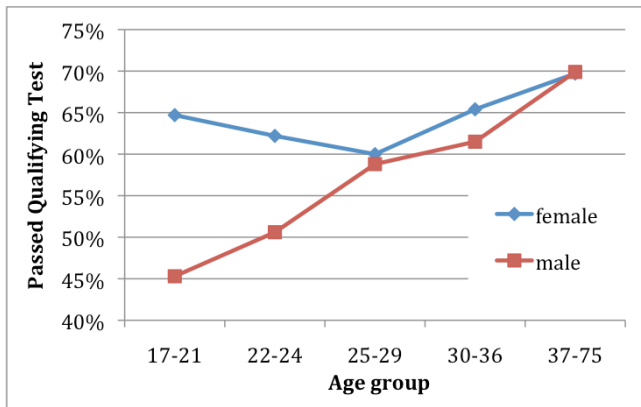
**Table 2. Occupations**

**Figure 1. Gender by Age Interaction**

and are not included in the demographic analyses. Forty-five respondents (2.3%) answered the demographics but neglected to answer both qualifying questions; their responses are included but considered incorrect. Reliability between the two questions was moderate, with 95% of those answering the difficult question correctly also answering the easy one correctly, $\chi^2=184.43$, $p=.001$.

Overall, 1,198 of the 1,962 participants (61%) qualified by answering both questions correctly. A total of 1,726 (88%) answered the easy question correctly, with similar numbers skipping (n=96, 5%) as getting it wrong (n=120, 6%). Only 1,266 (64%) answered the difficult question correctly, with an equivalent skip rate to the easier question (n=98, 5%) but far more respondents answering incorrectly (n=609, 31%).

Performance was examined as a function of gender (comparing men and women), age (broken into quintiles: 17-21, 22-24, 25-29, 30-36, 37-75), and 14 occupations grouped into 6 categories (see Table 2).

Women were more likely to answer the difficult question correctly than men (66% vs. 60%, $F(1,1884)=4.90$, $p<.05$). The same pattern held for the easy question, but was not significant (93% vs. 86%, $F(1,1884)=1.63$, $p=.20$).

Older participants were more likely to qualify than younger ones. Participants gave the correct answer more often as a linear function of age quintiles for both the difficult question, $F(1,1884)=13.35$, $p<.001$, and the easy one, $F(1,1884)=6.04$, $p=.014$.

There was a significant effect of reported occupational categories for the difficult question, $F(5,1884)=4.40$, $p<.001$ and a marginal trend for the easy question, $F(5,1884)=2.13$, $p=.059$. Professionals (69%) and students (71%) were more likely to answer the difficult question correctly compared to hourly workers (56%), financial workers (59%) and other occupations (60%). Those who were retired or not working were in between, not significantly different from any of the other groups (65%). A similar pattern emerged with the easy question, although fewer comparisons were significant perhaps due to ceiling effects. Professionals (91%), students (94%) and those not

working (92%) outperformed financial workers (85%), with hourly workers (88%) and unspecified (88%) not significantly different from any other group.

Qualification, requiring correct answers for both questions, revealed similar main effects for gender ($F(1,1884)=7.15$, $p<.01$; 64% of females and 57% of males qualified), age ($F(1,1884)=17.15$, $p<.001$), and occupation ($F(5,1884)=4.26$, $p<.001$). An interaction emerged between gender and age ($F(1,1884)=4.04$, $p<.05$) with performance as a linear effect of age for males ($p=.001$), ranging from 45% for the youngest quintile, age 17-21, to 70% for the oldest, age 37-75 (see Figure 1), but no effect of age for females ($p=.27$).

For the easy question alone, there was a slight trend toward an interaction between gender and occupation, $F(5,1884)=1.77$, $p=.12$. As Figure 2 shows, this appears to reflect very small gender differences for nonworking and financial workers, but large differences for hourly and other workers.

We explored the use of time stamps as a mechanism to identify participants who are clicking quickly rather than conscientiously to see if a simple assessment of time on task might predict performance on the qualifying questions. On average, participants spent 2 minutes (120 seconds) completing the task, ranging from 4 to 1,548 seconds, with a standard deviation of 119 seconds. Setting a speed threshold at the 90th percentile (45 seconds), we find that those who qualified were only slightly more likely to spend more than 45 seconds on the task (91%) compared to those who didn't qualify (88%, $\chi^2=4.69$, $p=.03$). Raising the threshold to the 95th percentile (28 seconds), doesn't much improve the differentiation (96% vs. 93%, $\chi^2=8.33$, $p<.01$). Lowering it to the 80th percentile (61 seconds) loses any ability to differentiate (80% vs. 79%, $\chi^2=0.21$, $p=.64$). Time on task was a better predictor of the easy question than of overall qualification, with only 76% of non-qualifiers taking longer than the 90th percentile threshold of 45 seconds ($\chi^2=33.90$, $p<.001$), but still the majority of those spending very little time on task would have qualified and thus appear to be answering the questions conscientiously.
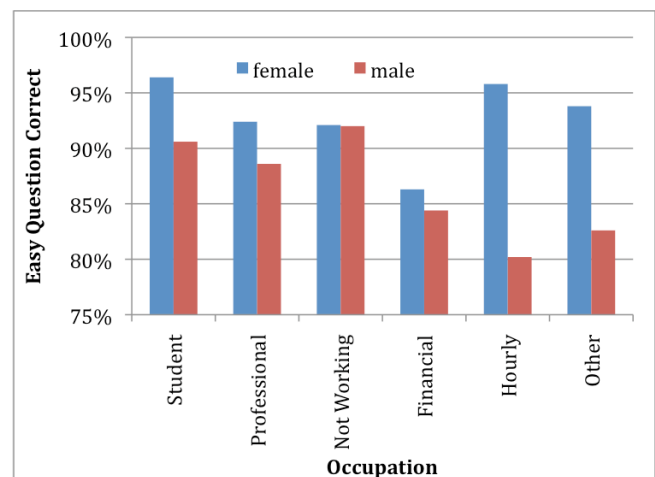


**Figure 2. Gender by Occupation Interaction**

Although time on task did not differentiate qualifiers from non-qualifiers very well, there was a mean difference between the groups, with non-qualifiers completing the task about 20 seconds more quickly than qualifiers, $t(1876)=3.39$, $p<.001$. Unfortunately, the considerable variance in both groups prevents time on task from being a reliable tool to differentiate these groups. We speculate that variability in computer load time and mouse maneuvering adds enough noise to overwhelm meaningful differences in cognitive processing time. Furthermore, people trying to game the system might not be lightening fast in their clicks, but rather could be acting distractedly, perhaps while doing something else simultaneously. In contrast, some conscientious participants may have quick computer response times and near-instant mousing or tabbing behavior. These data suggest that a threshold for time on task may not adequately identify non-conscientious participants, and may inadvertently disqualify many others.

## CONCLUSIONS

Some respondents may be participating in mTurk studies for quick cash rather than inherent interest, and may not be inclined to answer conscientiously. This screening tool provides a preliminary description of how people answer an easy and a difficult question, and who is likely to perform poorly. The easy question was a proxy for answering arbitrarily without even a cursory attempt to respond to content. The difficult question was a proxy for careful participation. No special knowledge or skill was required to answer either, just a willingness to do the task as presented. Getting either wrong is an indication that the participant may have been attempting to "game" the mTurk system.

Young men seem to be most likely to try to game the system, with fewer than half of men younger than 25 qualifying by getting both questions right. Men over 30 and women of any age were much more likely to qualify. Professionals, students, and non-workers seem to be the most likely to take the task seriously. It's possible that they tend to do mTurk tasks for the inherent interest and distraction, whereas hourly and financial workers may be trying to earn quick money while working at their normal jobs. Although the hourly rate offered by mTurk is small, if augmenting another income (e.g., administrative assistant waiting to be given a task to do) it can be a nontrivial source of additional money. The gender difference seems particularly strong among hourly workers, but cannot be explained by age or by differences between administrative and service jobs.

Interestingly, students tend to be relatively conscientious Turkers, especially given their younger age relative to other occupational categories (21.7 vs. 32.2, $t(1875)=21.82$, $p<.001$). Thus, a strategy to discourage younger people from participating might only need to focus on those who are not students.

## FUTURE WORK

Further piloting of similar kinds of qualifying questions would be informative. The usefulness of easy versus difficult questions has not been sufficiently explored by this study alone. Future work could attempt to create similar conditions that make lack of conscientiousness profitable in the lab, and determine whether performance on easy vs. hard questions is predictive of later performance. Additional studies could also identify parameters establishing optimal question design. One risk of moving forward with a cookie-cutter approach is that if many mTurk studies start using a similar, predictable design, then devious subjects might take to strategically answering screening questions, but not real survey questions. A system of embedding periodic screening questions could remedy that problem. Indeed, such an approach might have the optimal outcome of encouraging Turkers to complete tasks more conscientiously, rather than merely screening out those who don't.

## REFERENCES

1. Amazon Mechanical Turk. https://www.mturk.com

2. Amazon Mechanical Turk: Best Practices Guide. http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf

3. Atwood, J. Is Amazon's Mechanical Turk a Failure? *Coding Horror: Programming and Human Factors*, April 9, 2007. http://www.codinghorror.com/blog/archives/000828.html

4. Callison-Burch, C. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk, In *Proc. EMNLP 2009*, ACL and AFNLP (2009), 286–295.

5. Ipeirotis, P. Turker demographics vs. Internet demographics. http://behind-the-enemy-lines.blogspot.com/2009/03/turker-demographics-vs-internet.html. 2009.

6. Kittur, A., Chi, E.H., and Suh, B. Crowdsourcing User Studies with Mechanical Turk. In *Proc. CHI 2008*, ACM Press, (2008), 453-456.

7. MdFedries, P. Technically Speaking: It's a Wiki, Wiki World. *IEEE spectrum*, *43*, (2006), 88.

8. Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L., & Downs, J. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proc. CHI 2010*, ACM Press, (2010).