Moral Reminder as a Way to Improve Worker Performance on Amazon Mechanical Turk

Heeju Hwang

University of Hong Kong
heeju.hwang@gmail.com

Abstract

The present study explores a method to reduce abusive worker behavior on Amazon Mechanical Turk (AMT), namely reminding workers of moral standards. We manipulated workers' awareness of moral standards via the presence or the absence of an honesty statement in a survey. The results showed that the honesty statement significantly improved workers' performance during the first half of the survey. This suggests that a moral reminder is a simple and efficient way to reduce abusive worker behavior in a relatively short survey on AMT.

Introduction

Amazon Mechanical Turk (henceforth, AMT) is a crowdsourcing platform that provides researchers with an easy access to a large and diverse set of participants at a relatively low cost. For this reason, researchers from a variety of fields of social sciences, such as economics, psychology, and sociology have increasingly used AMT to conduct behavioral research. The use of AMT as a research platform, however, presents a challenge for assuring data quality. As workers will make most money by completing tasks as fast as possible, there is a potential for workers to abuse the system. That is, in order to decrease their time spent and thus increase their rate of pay, workers may provide invalid responses without attending to the instructions or purposes of tasks (e.g. Kittur, Chi, and Suh, 2008).

To evaluate the validity of AMT data, a number of studies compared AMT data with laboratory data. The results indicate that a small number of users took advantage of the system (e.g. Sprouse, 2011). Although a minority of users tend to generate invalid responses and their work can be rejected ex post, finding, removing and rejecting their responses consume experimenter resources and is prone to high measurement error (e.g. Paolocci and Chandler, 2014). Thus, the present study aims to address the question

of how to reduce invalid responses *beforehand* rather than screening them afterwards by reminding workers of moral standards.

Mazar, Amir, and Ariely (2008) found that when reminded of moral standards, participants were less likely to cheat, even if doing so reduced financial gains. In a series of experiments, participants were asked to solve problems and the amount of pay depended on the total number of correct answers. Mazar et al. provided participants with an opportunity to cheat by having them self-report the number of correctly solved problems. Crucially, before the task Mazar et al. manipulated participants' awareness of moral standards: In an experimental condition, participants were asked to sign a statement (i.e. honor code) in which they declared their commitment to honesty before engaging in the task (e.g. "I understand that this short survey falls under MIT's honor system."). In a control condition, such a moral reminder was not present. Mazar et al. found that when presented with the honor code, participants did not over-claim the number of problems solved correctly -- i.e. reminding participants of moral standards eliminated cheating. Mazar et al. suggest that a moral reminder affects dishonest behavior by influencing self-concept. More specifically, a moral reminder increases participants' awareness to moral standards and in order to maintain their self-concept as an honest individual, participants comply with the standards and behave honestly.

The results of Mazar et al. (2008) showed that reminding participants of moral standards was effective in eliminating cheating. Note, however, that the study concerns participants' (dis)honesty to self-report their performance, and does not address how a moral reminder influences participants' *performance* or behavior during the task -- i.e. whether participants complete the task in good faith or not.

Thus, the present study aims to investigate (i) whether a moral reminder influences the performance of a task, and (ii) if so, whether the effect of a moral reminder persists or decreases over the course of the task. For the purpose, we conducted a survey study on AMT. Following the methodology of Mazar et al. (2008), we manipulated participants'

awareness of moral standards via the presence or the absence of an honesty declaration before a survey (i.e. "I pledge to answer the following questions as accurately as possible."): In an experimental condition, participants were asked to read the honesty statement and type in their last initial below the statement before participating in the survey. In a control condition, participants were not presented with the honesty statement.

Method

Forty workers participated in the survey for $1.50. We restricted workers to be masters with approval rate greater than 95% and the location of their IP addresses to be from the United States. All workers indicated that English was their native language and they were over 18 years old. Each worker was randomly assigned to one of the two conditions.

The survey consisted of 24 sentence items. For each sentence, workers' task was to choose between two paraphrases whose meaning is consistent with the meaning of the given sentence as seen in (1).

(1) The boy who watched the housewife waxed the linoleum.
   a. The boy waxed the linoleum.
   b. The boy did not wax the linoleum.

We chose these types of sentences because accurate judgments to these sentences indicate participants' attentiveness to survey items and good faith (see Munro et al., 2010).

Participants' responses were categorized into correct and incorrect responses. In order to examine how the effect of a moral reminder varies over the course of survey, the survey items were categorized into two groups: the first half and the second half. The responses were fit using logit mixed-effects models. The final model includes the honesty statement, and the time course of survey and their interaction as fixed factors and random intercepts for participants and items.

Results and Discussion

Figure 1 plots per-worker percentage of correct responses when the honesty statement was absent and present during the first and the second half of the survey. As can be seen, the effect of the honesty statement was significant in the first half of the survey, but not in the second half. That is, compared to the performance of workers in the control condition (without the honesty statement), the performance of workers in the honesty statement condition was significantly better during the first half of the survey. A mixed-effects logistic regression model showed that there was a significant interaction effect between the honesty statement and the course of survey on the performance of workers

(β=0.1, z=2.360, SE=0.423, p<.01). The median completion time was greater in the honesty statement condition than in the control condition (340 seconds vs. 324 seconds), suggesting that participants in the honesty statement condition were likely to put more effort in the survey.
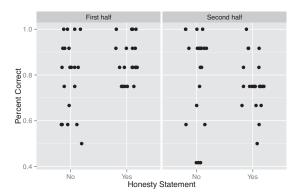

Figure 1. Per-worker Percentage of Correct Responses in the First and the Second Half of the Survey without and with the Honesty Statement.

The results suggest that reminding workers of moral standards is an effective method to improve worker performance for a relatively short survey task on AMT. The present study, however, is limited to an objective task (sentence comprehension survey) and does not inform us of whether a moral reminder is effective for a subject task such as rating. Thus, further research is necessary to generalize the findings beyond the context of the present study. Nonetheless, our findings suggest that at least for an objective task, a moral reminder is a simple cost-effective way to reduce abusive worker behavior, with potentially far-reaching implications for online survey designs.

References

Kittur, A., Chi, E. H., and Suh, B. 2008. Crowdsourcing user studies with Mechanical Turk. *Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, 453–456.

Mazar, N., Amir, O., and Ariely, D. 2008. The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, 45(6), 633-644.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., and Tily, H. 2010. Crowdsourcing and language studies: The new generation of linguistic data. *Proceedings of Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.* 122-130.

Paolacci, G., and Chandler, J. 2014. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, 23(3), 184-188.

Sprouse, J. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods,* 43(1), 155-167.