aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
6(1) 3–5

DOI: 10.1177/1745691610393980
http://pps.sagepub.com
SAGE

Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?

Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling
Department of Psychology, University of Texas at Austin

Abstract
Amazon's Mechanical Turk (MTurk) is a relatively new website that contains the major elements required to conduct research: an integrated participant compensation system; a large participant pool; and a streamlined process of study design, participant recruitment, and data collection. In this article, we describe and evaluate the potential contributions of MTurk to psychology and other social sciences. Findings indicate that (a) MTurk participants are slightly more demographically diverse than are standard Internet samples and are significantly more diverse than typical American college samples; (b) participation is affected by compensation rate and task length, but participants can still be recruited rapidly and inexpensively; (c) realistic compensation rates do not affect data quality; and (d) the data obtained are at least as reliable as those obtained via traditional methods. Overall, MTurk can be used to obtain high-quality data inexpensively and rapidly.

Keywords
Amazon Mechanical Turk, Internet, online, web, data collection, research methods

Amazon's Mechanical Turk (www.MTurk.com) is a novel, open online marketplace for getting work done by others. Here, we describe and evaluate the potential contributions that MTurk might make in psychology and other social sciences as a site for Web-based data-collection.

Introduction to MTurk

How Does MTurk Work?

MTurk functions as a one-stop shop for getting work done, bringing together the people and tools that enable task creation, labor recruitment, compensation, and data collection. The site boasts a large, diverse workforce consisting of over 100,000 users from over 100 countries who complete tens of thousands of tasks daily (Pontin, 2007). Individuals register as "requesters" (task creators) or "workers" (paid task completers). Requesters can create and post virtually any task that can be done at a computer (i.e., surveys, experiments, writing, etc.) using simple templates or technical scripts or linking workers to external online survey tools (e.g., SurveyMonkey). Workers can browse available tasks and are paid upon successful completion of each task. Requesters can refuse payment for subpar work. Being refused payment has negative consequences for workers because requesters can limit their tasks to workers with low refusal rates.

How Are Workers Compensated?

Requesters deposit money into an account using a credit card. Requesters set the compensation amount prior to posting a task; payments can be awarded automatically or manually based on the quality of each worker submission. Amazon charges a 10% commission.

Why Do Workers Participate?

Compensation in MTurk is monetary, but the amount awarded is typically small (e.g., nickels and dimes for 5-10 minute tasks). Our analyses (see online supporting materials at http://pps.sagepub.com/supplemental) of worker motivation suggest that they are internally motivated (e.g., for enjoyment).

Corresponding Author:
Michael Buhrmester, Department of Psychology, University of Texas at Austin, 1 University Station A8000, Austin, TX 78712
E-mail: buhrmester@gmail.com

Table 1. Effects of Compensation Amount and Task Length on Participation Rates (Submitted Surveys per Hour of Posting Time)

Compensation amount	Short survey (5 min)	Medium survey (10 min)	Long survey (30 min)
2 cents	5.6	5.6	5.3
10 cents	25.0	14.3	6.3
50 cents	40.5	31.6	16.7

Note. Surveys consisted of a series of demographic questions and personality scales. For the medium length survey, 60 participants were recruited per compensation amount. For the short and long surveys, 25 participants were recruited per compensation amount.

Evaluating the Quality of MTurk Data

How Do MTurk Samples Compare With Other Samples?

Commentators have long lamented the heavy reliance on American college samples in the field of psychology (Sears, 1986) and more generally those from a small sector of humanity (Henrich, Heine, & Norenzayan, 2010). Recent evidence suggests that collecting data via the Internet, although far from perfect, can reduce the biases found in traditional samples (Gosling, Vazire, Srivastava, & John, 2004).

To examine how MTurk samples compare with the diversity of standard Internet samples, we compared the demographics of 3,006 MTurk participants with those in a large Internet sample (Gosling et al., 2004). MTurk participants came from over 50 different countries and all 50 U.S. states. Gender splits were similar in the standard Internet (57% female) and MTurk (55% female) samples. A greater percentage of MTurk participants were non-White (36%) and almost equally non-American (31%) compared with the Internet sample (23% and 30%, respectively). MTurk participants were older ($M = 32.8$ years, $SD = 11.5$) than the Internet participants ($M = 24.3$ years, $SD = 10.0$). In short, MTurk participants were more demographically diverse than standard Internet samples and significantly more diverse than typical American college samples.

How Do Compensation Amount and Task Length Affect Participation Rates?

MTurk's major appeal is its potential for collecting data inexpensively and rapidly. To investigate participant response rates at various compensation levels and task lengths and to explore the tradeoffs between these parameters, we administered personality questionnaires via MTurk in a 3 × 3 design, crossing compensation level (2, 10, or 50 cents) with estimated task-completion time (5, 10, and 30 minutes).

There was a main effect of compensation level, $F(2, 6) = 20.67$, $p < .01$, with participation rates lowest in the 2-cent payment (see Table 1). With the exception of the 2-cent condition (due to a possible floor effect), there was a main effect of survey length such that response rates were lowest for the 30-minute survey, $F(1, 6) = 7.05$, $p < .05$. Note that although participation rates decreased as a function of both payment amount and survey length, we were still able to recruit participants for all conditions.

To explore the lower limits of compensation amount for task completion, we tested whether MTurk workers would complete a task for the lowest allowable payment rate: a penny. We posted a task that paid workers 1 cent for answering two pieces of information: age and gender. In 33 hours, we collected 500 responses or about 15 participants per hour. These results demonstrate that workers are willing to complete simple tasks for virtually no compensation, again suggesting that workers are not driven primarily by financial incentives.

These analyses suggest that participants can be recruited rapidly and inexpensively. Participation rates are sensitive to compensation amounts and time commitments, but our findings demonstrate that it is possible to collect decent-sized samples via MTurk for mere dollars. Even when offering just 2 cents for a 30-minute task, we accumulated 25 participants, albeit at a slower rate (i.e., in about 5 hours of posting time). Moreover, by increasing the compensation just slightly (e.g., to 50 cents) we were able to obtain the same number of participants in less than 2 hours of posting time.

How Does Compensation Amount Affect Data Quality?

To examine compensation-level effects on data quality, we computed alpha reliabilities for data collected at three levels of compensation (2, 10, and 50 cents) in a set of six personality questionnaires administered to MTurk participants. The mean alphas were within one hundredth of a point across the three compensation levels (see online supporting materials), suggesting that even at low compensation rates, payment levels do not appear to affect data quality; the only drawback appears to be data collection speed (as shown in the previous section), a finding consistent with previous research on nonsurvey tasks (Mason & Watts, 2009).

Do MTurk Data Meet Acceptable Psychometric Standards?

The absolute levels of the mean alphas were in the good to excellent range ($\alpha = .73$–$.93$; mean $\alpha = .87$ across all scales and compensation levels). Moreover, with three exceptions, the MTurk alphas were within two hundredths of a point of the traditional-sample alphas (see online supporting materials). To provide another index of data quality, we estimated test–retest reliabilities in a set of individual difference measures administered 3 weeks apart via MTurk. Participants

were paid 20 cents for completing Wave 1 and 50 cents for Wave 2 (60% completed them). Test–retest reliabilities were very high ($r = .80$–$.94$; mean $r = .88$) and compared favorably with test–retest correlations of traditional methods (see online supporting materials).

Summary and Conclusions

Our investigation into MTurk as a potential mechanism for conducting research in psychology and other social sciences yielded generally promising findings. The site has the necessary elements to successfully complete a research project from start to finish. Our analyses of demographic characteristics suggest that MTurk participants are at least as diverse and more representative of noncollege populations than those of typical Internet and traditional samples. Most important, we found that the quality of data provided by MTurk met or exceeded the psychometric standards associated with published research.

Still, the process of validating MTurk for use by researchers has only just begun. Some of MTurk's current strengths—the open market design and large, diverse participant pool—may change in the future (see online supporting materials for further discussion). That said, if future data continue to be as promising as they have proven here and elsewhere (e.g., Mason & Watts, 2009), we anticipate that MTurk will soon become a major tool for research in psychology and elsewhere in the social sciences.

Acknowledgment

We thank Matthew Brooks and William B. Swann, Jr. for feedback on an earlier version of this article.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

References

Gosling, S.D., Vazire, S., Srivastava, S., & John, O.P. (2004). Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, *59*, 93–104.

Henrich, J., Heine, S.J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 62–135.

Mason, W.A., & Watts, D.J. (2009). Financial incentives and the "performance of crowds." *Association for Computing Machinery Explorations Newsletter*, *11*(2):100–108.

Pontin, J. (2007, March 25). Artificial intelligence: With help from the humans. *The New York Times.* Retrieved from http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html

Sears, D.O. (1986). College sophomores in the lab: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515–530.