ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Current Directions in Psychological Science
2014, Vol. 23(3) 184–188
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0963721414531598
cdps.sagepub.com
SAGE

Inside the Turk: Understanding Mechanical Turk as a Participant Pool

Gabriele Paolacci[1] and Jesse Chandler[2,3]
[1]Department of Marketing Management, Rotterdam School of Management, Erasmus University Rotterdam; [2]Institute for Social Research, University of Michigan; and [3]PRIME Research, Ann Arbor, MI

Abstract

Mechanical Turk (MTurk), an online labor market created by Amazon, has recently become popular among social scientists as a source of survey and experimental data. The workers who populate this market have been assessed on dimensions that are universally relevant to understanding whether, why, and when they should be recruited as research participants. We discuss the characteristics of MTurk as a participant pool for psychology and other social sciences, highlighting the traits of the MTurk samples, why people become MTurk workers and research participants, and how data quality on MTurk compares to that from other pools and depends on controllable and uncontrollable factors.

Keywords
Mechanical Turk, online experimentation, data collection

The Internet has democratized knowledge by lowering barriers to the consumption, dissemination, and creation of knowledge. Although social scientists have long relied on the Internet for data collection, difficulties in recruiting and compensating participants have inhibited data collection online. A Web site called Mechanical Turk (MTurk) has recently offered a solution to these technical challenges.

MTurk is an online labor market created by Amazon to assist "requesters" in hiring and paying "workers" for the completion of computerized tasks. Tasks (e.g., transcribing text) are typically completed within minutes and usually pay in cents rather than dollars. Social scientists have recently discovered the potential of the MTurk workforce as a large pool of participants, constantly available to complete research studies at a low cost. Today, it is not uncommon to read empirical articles that are entirely based on data collected using MTurk.

With the surge of interest in MTurk as a participant-recruitment tool have come questions regarding its reliability. What are the characteristics of the MTurk population? Why do workers become research participants? Is the data collected on MTurk of adequate quality? Reservations are justified particularly because MTurk is *not* a participant pool, and it presents researchers with challenges that other pools do not (e.g., how to select participants on the basis of their characteristics). We integrate the available evidence that speaks to whether and how researchers can use MTurk as a data-collection tool.

Characteristics of MTurk Samples

Workers choose to complete MTurk tasks for minimal pay, which raises questions about who they are and why they do so. Although payment is an important factor, self-reports indicate that workers are driven by both extrinsic and intrinsic motives (e.g., workers have reported that they complete tasks "to make basic ends meet" and because "tasks are fun"; Paolacci, Chandler, & Ipeirotis, 2010; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010), which suggests that the rewards of working on MTurk are not merely monetary.

In 2014, the MTurk workforce is composed of more than 500,000 individuals from 190 countries. Demographic

Corresponding Author:
Gabriele Paolacci, Department of Marketing Management, Rotterdam School of Management, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, The Netherlands
E-mail: gpaolacci@rsm.nl

surveys consistently indicate that MTurk is dominated by workers residing in the United States and India, with less than a quarter of workers residing elsewhere (Paolacci et al., 2010; Ross et al., 2010). In general, workers are diverse but not representative of the populations they are drawn from, reflecting that Internet users differ systematically from non-Internet users. Workers tend to be younger (about 30 years old), overeducated, underemployed, less religious, and more liberal than the general population (Berinsky, Huber, & Lenz, 2012; Paolacci et al., 2010; Shapiro, Chandler, & Mueller, 2013). Within the United States, Asians are overrepresented and Blacks and Hispanics are underrepresented relative to the population as a whole (Berinsky et al., 2012).

Personality studies have also revealed differences between workers and other samples that parallel differences between frequent and infrequent Internet users (Shapiro et al., 2013). Worker samples are less extraverted than college-student and community samples (Goodman, Cryder, & Cheema, 2013; Kosara & Ziemkiewicz, 2010) and more socially anxious than the U.S. population at large (Shapiro et al., 2013). There is also some evidence that workers are less emotionally stable (Goodman et al., 2013; Kosara & Ziemkiewicz, 2010), although they are no more likely to display clinically relevant emotional disregulation than the general population (Shapiro et al., 2013).

Less is known about workers' cognitive abilities, and this remains a fruitful area of investigation. Paolacci and colleagues (2010) found no difference between workers, undergraduates, and other Internet users on a self-report measure of numeracy that correlates highly with actual quantitative abilities. However, workers may learn more slowly and have more difficulty with complex tasks than university students, perhaps reflecting differences in age and education (Crump, McDonnell, & Gureckis, 2013).

In sum, the pool of available workers is large and diverse. It can replace or supplement traditional convenience samples, but it should not be treated as representative of the general population. The sheer size of the MTurk workforce and the possibility to selectively recruit workers (described below) can also make it useful to reach samples with specific characteristics.

Sampling Issues on MTurk

Workers select the tasks they wish to perform from a large number of available alternatives. Importantly, workers are not necessarily qualified to complete all tasks. Requesters can decide who can and cannot participate in a task by requiring workers to possess "qualifications" that are assigned by MTurk (e.g., country of residence, ratio of approved to submitted tasks) or by requesters themselves (e.g., on the basis of information obtained in a prior study; J. Chandler, Mueller, & Paolacci, 2014). Researchers can thus intentionally or inadvertently construct samples that are not representative of the worker population.

How do workers choose among tasks for which they are qualified? By default, tasks are sorted according to how recently they were posted, and recency is a major determinant of participation (Chilton, Horton, Miller, & Azenkot, 2010). Workers can sort tasks according to additional criteria, including keywords and payments. Whereas few workers use keywords to find tasks (Chilton et al., 2010), more lucrative tasks are consistently more attractive (e.g., Buhrmester, Kwang, & Gosling, 2011; Mason & Watts, 2009). Interestingly, workers appear to be sensitive to nominal pay rates, finding tasks that pay in multiples of 5 cents to be more attractive (Horton & Chilton, 2010). The importance of recency and compensation means that participation peaks early and declines rapidly, and that researchers might acquire large samples more quickly by posting the task more than once or paying workers more.

Other factors affect the likelihood of workers' seeing and selecting a task and influence the sample composition, including task complexity (Kazai, Kamps, & Milic-Frayling, 2012), time of sampling (Komarov, Reinecke, & Gajos, 2013), and whether the task was discussed in an online forum (J. Chandler, Mueller, & Paolacci, 2014). Much remains to be learned about the factors that influence sample composition, but these findings underscore the need to collect and report sample characteristics rather than assume they are similar to those of earlier research studies, and caution against assuming that single studies will replicate within other MTurk samples.

The fact that workers are free to select what to work on and how much work to complete raises the concern that they participate in studies that employ procedures they have seen already. Researchers can be "followed" by workers who find their tasks lucrative and interesting, and the high rate at which researchers seem to post non-novel tasks on MTurk is unlikely to be matched by the rate at which new individuals join the workforce. We found evidence in our requester history that about 10% of workers are responsible for completing 41% of tasks, and that more experienced workers are more familiar with classic paradigms within behavioral sciences (e.g., trolley problems used in moral dilemmas: J. Chandler, Mueller, & Paolacci, 2014; see also Fort, Adda, & Cohen, 2011), which suggests that their prior experiences may influence their responses in research studies. Indeed, there is evidence that practice effects influence measures assumed to reflect individual differences. Worker productivity correlates with performance on a popular test of reflexivity but not with performance on a novel but logically identical test (J. Chandler, Mueller, & Paolacci, 2014), and Rand and colleagues (in press) found that workers' "intuitive" preference for cooperation declined with workers' experience but reemerged when assessed with a novel instrument. These effects can result from processing advantages gained by prior exposure to materials and the ability to draw on prior answers when forming a response. Importantly, practice effects do not generalize from specific to logically equivalent procedures, which suggests that these concerns can be avoided by developing and using novel research paradigms and excluding prior participants using qualifications or other methods (J. Chandler, Mueller, & Paolacci, 2014). If this is not possible, researchers should consider avoiding MTurk samples.

(Baseline) Data Quality

Although data quality can be defined in several ways, research assessing MTurk on dimensions universally relevant to researchers supports the idea that worker samples are reliable. Self-reports of individual differences are psychometrically valid (Buhrmester et al., 2011; Shapiro et al., 2013), and the quality of linguistic judgments respondents provide is comparable to that of college samples (Sprouse, 2011). Workers exhibit the same cognitive biases (e.g., framing effects), logical fallacies (e.g., conjunction fallacy), and behavior in economic games as traditional participants do (e.g., Amir, Rand, & Gal, 2012; Goodman et al., 2013; Horton, Rand, & Zeckhauser, 2011; Klein et al., 2014; Paolacci et al., 2010). Classic cognitive tasks that rely on response-time measures have also been replicated, including Stroop, switching, flanker, attentional blink, and subliminal-priming tasks (Crump et al., 2013). The primary limiting feature in these experiments is the technical quality of users' computers (i.e., the refresh rate of computer monitors) rather than noise induced by inattentive or distracted participants (Simcox & Fiez, 2014).

Monitoring Web participants is hard, and a natural concern is whether MTurk participants are conscientious and honest. In addition to the successful replication of attention-sensitive tasks, direct assessments of attentiveness revealed few differences between MTurk workers and other participants (e.g., Berinsky, Margolis, & Sances, in press; Paolacci et al., 2010). Workers also appear to be truthful when providing self-report information. Respondents' reported location typically matches their IP address (e.g., Rand, 2012; Shapiro et al., 2013), and there is remarkable consistency over time in workers' demographic characteristics (e.g., age and gender after 6 months: Mason & Suri, 2012) and individual-difference measures (after 1 week: Shapiro et al., 2013; and after 3 weeks: Buhrmester et al., 2011; Holden, Dennie, & Hicks, 2013). Indirect comparisons do not support suspicions that MTurk workers are more likely to cheat than members of college samples (Suri, Goldstein, & Mason, 2011), although the fact that workers do respond to incentives to cheat (Goodman et al., 2013) cautions against conducting studies on MTurk that provide participants with opportunities to exploit the uncontrollability of the online environment.

In sum, researchers can use MTurk for virtually any study that is feasible to conduct online. Workers are diligent because of their intrinsic motivations and the incentive structure of MTurk: Requesters are not forced to approve submissions and can screen workers on the basis of past approval rates.

Although the desire to provide quality responses is usually beneficial, it may also lead to unintended consequences. MTurk workers seem to score higher in social desirability (Behrend, Sharek, Meade, & Wiebe, 2011), are unusually likely to report rare symptoms that appear clinically relevant (Shapiro et al., 2013), and may search the Internet for the answers to factually verifiable questions (Goodman et al., 2013). This suggests that workers are motivated to please requesters, and it highlights the importance of taking measures to avoid demand effects by concealing the research question of interest, temporally separating prescreening measures from the dependent variables of interest, and using between-subjects experimental manipulations.

Determinants of Data Quality

Data quality on MTurk is good, but it is also variable. Intuitively, paying workers more should motivate them. Indeed, payment improves performance on tasks with factually correct answers that can be determined through additional effort (Aker, El-Haj, Albakour, & Kruschwitz, 2012) and reduces random responses (Kazai, 2010). However, for tasks that rely on subjective responses, as most psychology studies do, there is no relationship between pay rates and quality (Buhrmester et al., 2011; Marge, Banerjee, & Rudnicky, 2010; Mason & Watts, 2009). In these cases, true responses likely require no more effort than false responses, and the limiting factor is more likely to be workers' understanding of the task (e.g., language comprehension; Goodman et al., 2013), which cannot be reduced by paying more.

Researchers who use MTurk often "improve" data quality by using screening methods that exclude problematic observations ex post (J. Chandler, Mueller, & Paolacci, 2014). There are multiple reasons to avoid using ex post screening methods, particularly those that flag inattentive participants using one or a small number of questions. These have high measurement error, rely on the questionable assumption that measured attentiveness is constant throughout the task, and may tap into correlated traits rather than state-level differences in attentiveness (Berinsky et al., in press), sacrificing sample diversity. There is also no evidence that attention checks improve data quality above and beyond simply recruiting workers with a high approval rate (Peer, Vosgerau, & Acquisti, 2013).

Attention may vary even within conscientious workers, and researchers may want to induce attentiveness ex ante by including tasks that signal to workers that their subsequent responses will be scrutinized (e.g., tasks in which participants must answer factually verifiable questions; Heer & Bostock, 2010; Kittur, Chi, & Suh, 2008) or require commitment (Rand, Greene, & Nowak, 2012). Performance-contingent bonuses increase passing rates on factual manipulation checks, perhaps by signaling that certain information should be attended to (J. Chandler, Paolacci, & Mueller, 2014). Moreover, consistent with the importance of intrinsic motivation among workers, data quality can be increased by embedding tasks with "meaning." Thanking workers and explaining to them the meaning of the task they will complete can stimulate better work (D. Chandler & Kapelner, 2013), as does framing a task as requested by a nonprofit organization (Rogstadius et al., 2011).

Conclusions

MTurk has accelerated and democratized science by facilitating access to a heterogeneous research-participant pool and has provided scientists with a platform to conduct research that is hard to conduct within physical labs or elsewhere online, such as studies of the real-time dynamics of large groups (e.g., Mason & Watts, 2012), cross-cultural comparisons (between the United States and India; Eriksson & Simpson, 2010), longitudinal studies (e.g., Berinsky et al., 2012), and field experiments (D. Chandler & Kapelner, 2013).

These technical advantages are compelling, but researchers should use MTurk with care. While more diverse than college samples, workers (like Internet users) are not a representative sample, and sample composition varies dynamically. The potential for arbitrary design choices to influence sample composition suggests that researchers should be transparent in the materials used in their studies and the methods used to recruit and exclude participants. The low cost of MTurk data facilitates the collection of well-powered samples that, ceteris paribus, better reflect the available workforce. Further, tools exist to recruit desired workers (e.g., those who are attentive and produce quality work) and avoid undesired workers (e.g., those who are known to have participated in similar studies in the past), allowing further control over the final sample. Finally, important questions remain unanswered about the cognitive profile of workers at large; the specifics of how prior experience, community norms, and other factors influence survey response; and how sampling decisions (e.g., when the task is posted and how it is described) influence the characteristics of MTurk samples.

Recommended Reading

Chandler, J., Mueller, P., & Paolacci, G. (2014). (See References). An article that discusses and explores empirically several methodological concerns connected to MTurk experimentation.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). (See References). An article that, in addition to validating MTurk for experimental research in economics, discusses extensively issues related to experimental validity on MTurk.

Mason, W., & Suri, S. (2012). (See References). An article that provides an overview of MTurk and can be a useful beginner's guide for researchers considering collecting data from MTurk workers.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). (See References). An article that replicates classic findings in the decision-making literature and benchmarks MTurk against other participant pools.

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

Aker, A., El-Haj, M., Albakour, M.-D., & Kruschwitz, U. (2012, May). *Assessing crowdsourcing quality through objective tasks*. Paper presented at the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey.

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the Internet: The effect of $1 stakes. *PLoS ONE*, *7*(2), e31461.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*, 800–813.

Berinsky, A., Margolis, M., & Sances, M. (in press). Separating the shirkers from the workers? Making sure respondents pay attention on Internet surveys. *American Journal of Political Science*.

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

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*, 3–5.

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, *90*, 123–133.

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.

Chandler, J., Paolacci, G., & Mueller, P. (2014). Risks and rewards of crowdsourcing marketplaces. In P. Michelucci (Ed.), *Handbook of human computation* (pp. 377–392). New York, NY: Springer.

Chilton, L. B., Horton, J. J., Miller, R. C., & Azenkot, S. (2010, July). *Task search in a human computation market*. Paper presented at the Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining Workshop on Human Computation, Washington, DC.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), Article e57410. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0057410

Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, *5*, 159–163.

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, *37*, 413–420.

Goodman, J. K., Cryder, C. E., & 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.

Heer, J., & Bostock, M. (2010, April). *Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design*. Paper presented at the Special Interest Group on Computer-Human Interaction Conference on Human Factors in Computing Systems, Atlanta, GA.

Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5-120 on Amazon's Mechanical Turk. *Computers in Human Behavior*, *29*, 1749–1754.

Horton, J. J., & Chilton, L. B. (2010, June). *The labor economics of paid crowdsourcing*. Paper presented at the 11th Association for Computing Machinery Conference on Electronic Commerce, Cambridge, MA.

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

Kazai, G. (2010, October). *An exploration of the influence that task parameters have on the performance of crowds*. Paper presented at CrowdConf 2010, San Francisco, CA.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2012, October–November). *The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy*. Paper presented at the Proceedings of the 21st Association for Computing Machinery International Conference on Information and Knowledge Management, Lahaina, HI.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In M. Burnett, M. F. Costabile, T. Catarci, B. de Ruyter, D. Tan, M. Czerwinski, & A. Lund (Eds.), *Chi 2008: 26th Annual Chi Conference on Human Factors in Computing Systems Conference Proceedings* (Vol. 1, pp. 453–456).

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . & Nosek, B. A. (in press). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*.

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013, April–May). *Crowdsourcing performance evaluations of user interfaces*. Paper presented at the Proceedings of the Special Interest Group on Computer-Human Interaction Conference on Human Factors in Computing Systems, Paris, France.

Kosara, R., & Ziemkiewicz, C. (2010, April). *Do Mechanical Turks dream of square pie charts?* Paper presented at the Proceedings of the 3rd BELIV'10 Workshop: Beyond Time and Errors: Novel Evaluation Methods for Information Visualization, Atlanta, GA.

Marge, M., Banerjee, S., & Rudnicky, A. I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 5270–5273). Retrieved from http://dblp.uni-trier.de/db/conf/icassp/icassp2010.html#MargeBR10

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1–23.

Mason, W., & Watts, D. J. (2009, June). *Financial incentives and the performance of crowds*. Proceedings of the Human Computation Workshop: Knowledge Discovery and Data Mining Conference, Paris, France.

Mason, W., & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences, USA*, *109*, 764–769.

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

Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-013-0434-y

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.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*, 427–430.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (in press). Intuitive cooperation and the social heuristics hypothesis: Evidence from 15 time constraint studies. *Nature Communications*.

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., & Vukovic, M. (2011, July). *An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets*. Paper presented at the 5th International AAAI Conference on Weblogs and Social Media, Barcelona, Spain.

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010, April). *Who are the crowdworkers?: Shifting demographics in Mechanical Turk*. Paper presented at the CHI'10 Extended Abstracts on Human Factors in Computing Systems, Atlanta, GA.

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

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods*, *46*, 95–111.

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

Suri, S., Goldstein, D. G., & Mason, W. A. (2011, August). *Honesty in an online labor market*. Paper presented at the 3rd Association for the Advancement of Artificial Intelligence Human Computation Workshop, San Francisco, CA.