An Evaluation of Amazon's Mechanical Turk, its Rapid Rise, and Effective Use

Journal:	*Perspectives on Psychological Science*
Manuscript ID	PPS-17-079.R1
Manuscript Type:	Special Issues Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Buhrmester, Michael; University of Oxford, Institute of Cognitive and Evolutionary Anthropology Talaifar, Sanaz; University of Texas, Austin, Psychology Gosling, Samuel; University of Texas, Austin, Psychology
Keywords:	Methodology: Scientific, Methodology: Quantitative
User Defined Keywords:	Internet Research, Online methods, MTurk

An Evaluation of Amazon's Mechanical Turk, its Rapid Rise, and Effective Use

Michael D. Buhrmester

University of Oxford

Sanaz Talaifar

University of Texas at Austin

Samuel D. Gosling

University of Texas at Austin

University of Melbourne

Correspondence concerning this article should be addressed to Michael D. Buhrmester at

Buhrmester@gmail.com

Abstract

Over the past two decades, many social scientists have expanded their data-collection capabilities by utilizing various online research tools. The 2011 article, *Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?* (Buhrmester, Kwang, & Gosling, 2011), introduced researchers to what was then considered to be a promising but nascent research platform. Since then, thousands of social scientists from seemingly every field have conducted research using the platform. Here, we reflect on the impact of Mechanical Turk on the social sciences and our article's role in its rise, provide the newest, data-driven recommendations to help researchers effectively utilize the platform, and highlight other online research platforms worth consideration.

Keywords: Online methods, Internet research, MTurk

An Evaluation of Amazon's Mechanical Turk, its Rapid Rise, and Effective Use

In 2009, when we began working on our original article, "*Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?",* the phrase "Amazon's Mechanical Turk" (or "MTurk" for short) sounded like word salad to most social scientists. Times have changed. In less than a decade, thousands of researchers across the social sciences have conducted research using MTurk. How and why did this change occur? And what does the future hold for MTurk and the next generation of online research platforms? Here, we shed some light on these questions as they relate to our original article.

A brief summary of our original article

Our article had three aims: 1) introduce MTurk to unfamiliar readers, 2) evaluate its utility for conducting academic research, and 3) encourage continued evaluation of MTurk. We found that after mastering the basics, conducting research using MTurk could be efficient and relatively inexpensive. Most important, we found that MTurk participants provided data that met or exceeded the psychometric standards set by data collected using other means (e.g., undergraduate samples). Thus, we concluded that the platform may serve as a useful tool for many social scientists, and we encouraged others to continue to evaluate MTurk because only a few such evaluations existed at the time (e.g., Mason & Watts, 2009; Paolacci, Chandler, & Ipeirotis, 2010, Ross et al., 2010).

The genesis of our article and its impact

In late 2009, discovering that no one at our weekly area meeting had heard of MTurk except for us, we decided to explore the platform and relevant literature. Initially, we were suspicious that MTurk was too good to be true, so we set out to evaluate several fundamental questions about the platform. The results were mostly encouraging, so we thought it would be helpful to introduce MTurk to a wider audience and encourage continued evaluation of the platform.

The article quickly became highly cited (4265 cites as of March 2017 according to Google Scholar). Year over year, citations of the article have increased, from 61 cites in 2011 to 1146 in 2016, suggesting that more and more researchers are finding the article useful. We hoped our article would interest a broad range of social scientists, and it appears we were successful in that regard; articles that cited ours appeared in over 1000 different journals. Our article also succeeded in encouraging others to continue evaluating the suitability of MTurk for conducting research. The literature since 2011 includes over 40 articles that have conducted some form of evaluation of MTurk's suitability. However, it is difficult to determine exactly how much impact our article had on the resulting explosion of research utilizing and evaluating MTurk because a number of other similar articles came out at about the same time (e.g., Berinsky, Huber, & Lenz, 2012, Mason & Watts, 2009, Paolacci et al., 2010), most of which are also highly cited.

MTurk's impact

Regardless of the unique effect of our original article, it is clear that MTurk itself has had a significant impact on the social sciences. Some quick facts: In social science journals with an impact factor greater than 2.5, 2011 saw fewer than 50 papers using data from MTurk whereas 2015 saw more than 500 (Chandler & Shapiro, 2016). In 2012, less than 10% of papers appearing in *JPSP*, *PSPB,* and *Psychological Science* contained at least one MTurk study; in 2015, between 20%-45% contained at least one (Zhou & Fishbach, 2016). Instead of using MTurk merely for conducting simple, cross-sectional surveys as some had feared, researchers have utilized multiple methodologies. In fact, by our count, over the past three years more than half of MTurk studies reported in *JPSP* utilized experimental methods.

For many researchers, our article was likely their first introduction to MTurk and led them to learn more about the platform's utility. The rapid and widespread growth in use of MTurk was also facilitated by the confluence of other developments. Perhaps the largest

development was social scientists' increasing acceptance of and eagerness to conduct Internet-based research, changes that began to take place in the late 1990s and 2000s (Gosling & Mason, 2015). During this period, myths about online participants as unmotivated, social misfits were empirically dispelled (Gosling, Vazire, Srivastava, & John, 2004), companies like Millisecond and Qualtrics began to offer increasingly sophisticated methodological tools online, and an increasing proportion of human behavior began to take place online (Van Dijck, 2013). And although scores of useful texts detailed how to conduct online projects before MTurk existed (e.g., Fraley, 2004; Gosling & Johnson, 2010), MTurk offered a relatively shallow learning curve compared to other online methods. Many researchers have also created guides tailored to fellow researchers (e.g., Buhrmester's online MTurk Guide for Social Scientists; Mason & Suri, 2012; Sheehan & Pittman, 2016).

How can MTurk continue to be useful?

As with any means of collecting data, MTurk has strengths and limitations to consider, ethical issues to navigate, and practical hurdles to overcome. Fortunately, researchers from across the social sciences have increasingly engaged in evaluating these issues. On balance, these evaluations show that MTurk continues to be useful across many research contexts as long as researchers consider the issues raised by these evaluations and employ best practices.

It is beyond the scope of this article to review all the relevant evaluations in depth. Instead, we highlight the major foci of these evaluations. We have organized them as a set of questions for getting the most out of MTurk. For more detailed analysis, we point readers to the recent, comprehensive review by Chandler and Shapiro (2016). Many of our considerations echo their points.

Can your study be practically implemented via MTurk?

Range of uses. In most cases, if a study can be conducted via the Internet, it can be

conducted via MTurk. Some designs can be executed entirely within MTurk's system, while other designs require pairing MTurk with other online platforms (Chandler & Shapiro, 2016). Learning about MTurk or other online platforms may itself spark new and creative designs (e.g., collaborative complex problem solving tasks; Mason & Watts, 2011). Not all designs, however, are well-suited for MTurk (e.g., extremely time-intensive surveys, tasks that require a totally distraction-free environment).

The participant pool. One feature of MTurk is the ability to collect large samples that are often more demographically diverse than typical undergraduate populations (Buhrmester et al., 2011; Casler, Bickel, & Hackett, 2013). In recent years, researchers have estimated that roughly 7,000 active workers comprise the participant pool. On average, the pool turns over every seven months and is composed of mostly Americans (Stewart et al., 2015). The pool, however, is not representative of the U.S. population (Arditte, Cek, Shaw & Timpano, 2016; see MTurk Tracker, 2016 for up-to-date pool characteristics), and data quality can be variable from participants for whom English is a second language (Goodman, Cryder, & Cheema, 2013). Due to the pool's size and turnover, researchers may be able to find samples of hard-to-reach populations, such as participants with rare mental or physical health symptoms (Chandler & Shapiro, 2016, Gillan & Daw, 2016). However, when recruiting rare populations it is essential to use prescreening measures (e.g., masking qualification criteria, preventing duplicate responding, etc.) to prevent fraudulent responses (Chandler & Paolacci, in press).

Speed, cost, and accessibility. Data can usually be collected rapidly on MTurk, but payment amount, time to complete the study, the size of the target population, and other factors influence response rates (Buhrmester et al., 2011) and data quality (Litman, Robinson, & Rosenzweig, 2015). MTurk's cost can be lower than many other methods, such as paying community members to participate in lab studies or paying for online platforms like Qualtrics

to collect data from their participant panels. Creating an MTurk account is easy but limited to certain countries (e.g., U.S., U.K.). For many researchers who cannot gain efficient access to participants (e.g., because their university has no participant pool), MTurk may represent the only efficient option for collecting data.

Are you minimizing factors that negatively affect data quality?

Inattention. Some evaluations have found that MTurk participants' attention is equal to or better than undergraduate participants' attention (Chandler & Shapiro, 2016; Hauser & Schwarz, 2015). However, as in student populations, varying levels of inattention can occur. As a first line of defense against inattention, ensure that instructions are clear and intuitive (Ramsey, Thompson, McKenzie, & Rosenbaum, 2016). Attention check questions may help quantify inattention levels and provide a rationale for discarding data (e.g., Oppenheimer, Meyvis, & Davidenko, 2009), but they also have downsides. They do not guarantee increased attention, may heighten attrition (Berinsky, Margolis, & Sances, 2016), and change how participants respond to critical thinking tasks (Hauser & Schwarz, 2015). Compared to attention checks, restricting participation to participants with a 95% approval rate or higher is equally effective at reducing inattention, thus leading some experts to generally recommend against the use of attention checks (Peer et al., 2014).

Non-naïveté and dishonesty. Non-naïve participants may compromise data quality. To deter non-naïveté, the MTurk system disallows participants from requesting payment for the same HIT more than once. However, MTurk does not prohibit participants from completing similar studies or commonly used stimuli (e.g., the trolley dilemma) more than once. Thus, researchers should take steps to identify and prohibit non-naïve participants from their studies. Simple prescreening questions can be effective, as well as utilizing MTurk's customizable qualification system (Chandler, Mueller, & Paolacci, 2014). In addition, cross-talk – discussing the study with other potential participants while the study is still being

conducted – is possible on MTurk but rarely occurs (Chandler et al., 2014).

Lastly, although the online nature of participation enables high levels of anonymity, thus in theory promoting honesty, dishonest responding can still occur (Rand, 2012), especially when participants suspect that dishonest answers will allow them to meet study inclusion criteria (Chandler & Paolacci, in press). Rates of dishonest responding on MTurk vary greatly, from near zero dishonesty on some general knowledge questions (Clifford & Jerit, 2016) to high levels of dishonesty in studies where participants lie about themselves to meet known study inclusion criteria (Chandler & Paolacci, in press). Explicitly encouraging honesty can effectively reduce the problem (Lowry, D'Arcy, Hammer, & Moody, 2016; Clifford & Jerit, 2016) as can certain prescreening measures (Chandler & Paolacci, in press).

Attrition. Given the relative ease with which participants can withdraw from online studies as compared to lab studies, researchers should be especially attentive to the possibility of systematic attrition on MTurk (Horton, Rand, & Zeckhauser, 2011; Zhou & Fishbach, 2016). Multiple steps can be taken to minimize attrition, including common-sense steps such as accurately estimating the time it will take participants to complete the study so they can plan accordingly (Chandler & Shapiro, 2016). However, even when careful steps are taken to minimize data-quality issues, they can still arise. Several authors suggest creating a priori rules about data inclusion/exclusion, carefully tracking attrition rates, and understanding how attrition can influence study results as well as the types of analyses that should be conducted (Mason & Suri, 2012; Zhou & Fishbach, 2016).

Are you following steps to ensure that participants are treated ethically?

Researchers should familiarize themselves with the ethics involved in online research (Buchanan & Williams, 2010) and ethics unique to MTurk. For instance, there are unequal power dynamics between requesters and workers on MTurk, and fair pay is a common

concern (Chandler & Shapiro, 2016; Gleibs, 2016). We also recommend that researchers spend time as a worker on the platform to get a better sense of participants' experiences and the ethical issues involved.

Are you fully reporting how MTurk was utilized?

Under-reporting of methods has been a serious concern across the social sciences, and this problem applies to MTurk as well. Researchers often neglect to mention many details such as attrition rates in each condition, measures of participant non-naïveté or inattention, and whether any of MTurk's default qualifications were used (e.g., worker approval rating %, geographical restrictions, etc.). These methodological decisions can have important effects on sample characteristics and, ultimately, study results. Transparency is key –report all of the relevant analysis and design considerations above.

Beyond MTurk

MTurk is just one of a growing number of online research platforms available to researchers. For example, Prolific Academic, founded in 2014 specifically for academic research, is similar to MTurk but with a more diverse participant pool (Peer, Brandimarte, Samate, & Acquisti, 2017). Turkprime similarly caters to academic users. Several other companies, commonly used for creating online surveys and experiments (SurveyMonkey, Qualtrics, and SurveyGizmo), now also offer access, for a sizeable fee, to online panels that include substantial numbers of participants outside the U.S. In addition, there are many established and new online platforms geared primarily toward market research but which are potentially also suitable for other types of research (e.g., Knowledge Networks, Research Now, Toluna; Callegaro et al., 2014). Many of these services claim to achieve high levels of representativeness, albeit usually at significant cost. Each of the platforms above has been used by at least a handful of academic researchers whose findings are reported elsewhere

(e.g., Black & Reynolds, 2012; Rizvi & Bobocel, 2016). Thus, before heading into seemingly uncharted waters, it would be prudent to seek out these pioneers and learn from their experiences.

In conclusion, based on the findings of a large number of evaluations over the past few years, on balance, MTurk remains a useful method for conducting a wide range of research. Its utility, however, is dependent on employing best practices and carefully considering the issues raised by MTurk's many evaluators. As in our original article, we encourage researchers to continue to evaluate MTurk empirically and in comparison to traditional offline methods and emerging online platforms. In addition, we encourage researchers to explore the challenging, big-picture questions about the lasting impact of MTurk and other online platforms. For instance, have such platforms generally aided researchers who have faced data collection challenges in the past? And have they had a positive or negative impact on the tempo of data collection and generation of new scientific knowledge?

We are in the midst of a technological revolution in the social sciences that extends beyond MTurk and its ilk to new methods such as mobile sensing (Harari et al., 2017), scraping data from social media (Kosinski, Matz, Gosling, Popov, & Stillwell, 2015), and global recruitment reach using sites like Google AdWords (e.g., Antoun, Zhang, Conrad, & Schober, 2016). The rapid rise of MTurk is just one sign of the changes that are underway in social scientific research. Thus, careful, consistent, and coordinated evaluation is more important than ever.

Acknowledgments

We thank Rachel Krakauer and Aaron Wood for assisting in the organization of literature and Matt Brooks for comments.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest.

References

Antoun, C., Zhang, C., Conrad, F. G., & Schober, M. F. (2016). Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field Methods*, *28*(3), 231-246.

Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment*, *28*(6), 684.

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.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2016). Can we turn shirkers into workers?. *Journal of Experimental Social Psychology*, *60*, 20-28.

Black, J., & Reynolds, W. M. (2013). Examining the relationship of perfectionism, depression, and optimism: Testing for mediation and moderation. *Personality and Individual Differences, 54*(3), 426-431.

Buchanan T., &Williams J.E. (2010). Ethical issues in psychological research on the Internet. In Gosling, S. & Johnson, J., eds. *Advanced Methods for Conducting Online Behavioral Research.* pp. 255–71.

Buhrmester, M.D., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5.

Callegaro, M., Baker, R. P., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (Eds.). (2014). *Online panel research: A data quality perspective*. John Wiley.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156-2160.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112-130.

Chandler, J. & Paolacci, G. (in press). Lie for a Dime: When most prescreening responses are honest but most study participants are imposters. *Social Psychological and Personality Science*.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Clinical Psychology*, *12*(1), 53.

Fraley, R. C. (2004). How to conduct behavioral research over the internet: A beginner's guide to HTML and CGI/Perl. New York: Guilford.

Gillan, C. M., & Daw, N.D. (2016). Taking psychiatry research online. *Neuron, 91*(1), 19-23.

Gleibs, I. H. (2016). Are all "research fields" equal? Rethinking practice for the use of data from crowdsourcing market places. *Behavior Research Methods*, 1-10.

Gosling, S. D., & Johnson, J. A. (2010). *Advanced methods for conducting online behavioral research*. American Psychological Association.

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. *Annual Review of Psychology*, *66*, 877-902.

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*(2), 93.

Hauser, D. J. & Schwarz, N. (2015). IT'S A TRAP!: Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open*, *5*(2), 1-6.

Harari, G. M., Gosling, S. D., Wang, R., Chen, F., Chen, Z. & Campbell, A. T. (2017). Patterns of behavior change in students over an academic term: a preliminary study of activity and Sociability Behaviors Using Smartphone Sensing Methods. *Computers in Human Behavior*, *67*, 129-138.

Hauser, D., & Schwarz, N. (2015). Attentive Turkers: MTurk participants perform better on online attention checks than subject pool participants. *Behavior Research Methods*.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399-425.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist, 70,* 543-556.

Litman, L., Robinson, J. & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, *47*(2), 519-528.

Lowry, P. B., D'Arcy, J., Hammer, B., & Moody, G. D. (2016). "Cargo Cult" science in traditional organization and information systems survey research: A case for using nontraditional methods of data collection, including Mechanical Turk and online panels. *The Journal of Strategic Information Systems*, *25*(3), 232-240.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical

Turk. *Behavior Research Methods*, *44*(1), 1-23.

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

Mason, W., & Watts, D. J. (2011). Collective problem solving in networks. *SSRN 1795224*.

Mechanical Turk Monitor (2016). http://www.mturk-tracker.com.

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

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023-1031.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163.

Ramsey, S. R., Thompson, K. L., McKenzie, M., & Rosenbaum, A. (2016). Psychological research in the internet age: The quality of web-based data. *Computers in Human Behavior*, *58*, 354-360.

Rizvi, S., & Bobocel, D. R. (2016). Promoting Forgiveness Through Psychological Distance. *Social Psychological and Personality Science*, *7*(8), 875-883.

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in mechanical turk. In *CHI'10 extended abstracts on human factors in computing systems* (pp. 2863-2872). ACM.

Sheehan, K.B., & Pittman, M. (2016). Amazon’s Mechanical Turk for Academics. Melvin & Leigh Publishers.

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*(5), 479.

Van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.

Zhou, H., Fishbach, A., Shaddy, F., Steinmetz, J., Bregant, J., Schroeder, J., ... & Choshen-Hillel, S. (2016). The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (yet False) Research Conclusions. *Journal of Personality and Social Psychology, forthcoming*.