Kardes, F. R. (1996). In defense of experimental consumer psychology. *Journal of Consumer Psychology, 5*, 279–296.
Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizations, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice*.
Locke, E. A. (1986). Generalizing from laboratory to field: Ecological validity or abstraction of essential elements? In E. A. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 257–267). Lexington, MA: Heath.
Mitchell, G. (2012). Revisiting truth or triviality the external validity of research in the psychological laboratory. *Perspectives on Psychological Science, 7*(2), 109–117.
Mook, D. (1983). In defense of external invalidity. *American Psychologist, 38*, 379–387.
Oakes, W. (1972). External validity and the use of real people as subjects. *American Psychologist, 27*, 959–962.
Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*, 184–188.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411–419.
Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*(3), 515.
Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology, 96*(5), 1055–1064.
Stone-Romero, E. F. (2002). The relative validity and usefulness of various empirical research designs. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 77–98). Malden, MA: Blackwell.

Caution! MTurk Workers Ahead–Fines Doubled

P. D. Harms
University of Alabama

Justin A. DeSimone
University of Cincinnati

Landers and Behrend (2015) are the most recent in a long line of researchers who have suggested that online samples generated from sources such as Amazon's Mechanical Turk (MTurk) are as good as or potentially even better than the typical samples found in psychology studies. It is important that the

P. D. Harms, Department of Management, University of Alabama; Justin A. DeSimone, Organizational Leadership and Human Resources, University Cincinnati.

Correspondence concerning this article should be addressed to P. D. Harms, University of Alabama, 101 Alston Hall, Box 870225, 361 Stadium Drive, Tuscaloosa, AL 35487. E-mail: pharms@gmail.com

https://doi.org/10.1017/iop.2015.23 Published online by Cambridge University Press

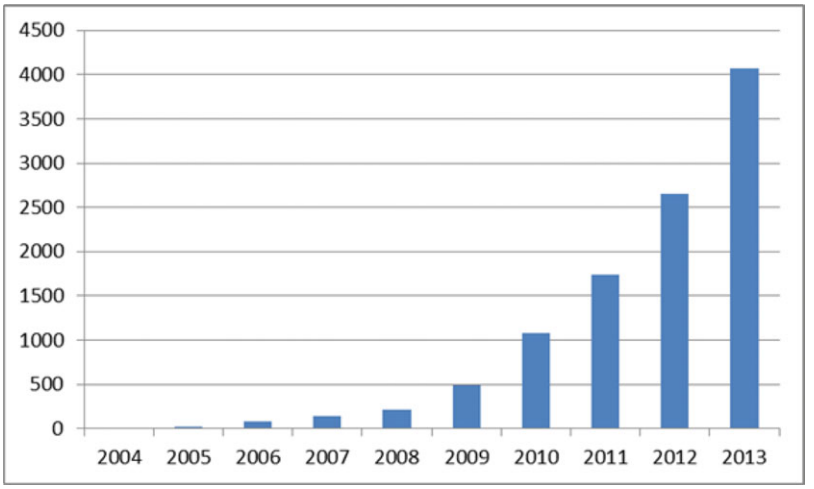


Figure 1 Google Scholar Search of "Mechanical Turk" (Color online)

authors caution that researchers and reviewers need to carefully reflect on the goals of research when evaluating the appropriateness of samples. However, although they argue that certain types of samples should not be dismissed out of hand, they note that there is only scant evidence demonstrating that online sources can provide usable data for organizational research and that there is a need for further research evaluating the validity of these new sources of data. Because the target article does not directly address the potential problems with such samples, we will review what is known about collecting online data (with a particular focus on MTurk) and illustrate some potential problems using data derived from such sources.

Although there have only been a few articles to date using MTurk samples in premiere organizational journals, the usage of MTurk samples both inside and outside the organizational sciences is increasing. Google Scholar searches for "mechanical turk" demonstrate its rapid ascendance in the academic literature (see Figure 1). Searching within those results for the terms "job performance," "job satisfaction," and "leadership" shows a similar pattern for topics that are of interest to organizational scholars (see Figure 2). Given the rapidly increasing interest in online samples as data sources for organizational researchers, it is clear that there is a pressing need to come to a better understanding of such samples.

One of the most widely studied topics in research methods papers investigating MTurk and other online samples concerns their representativeness (e.g., Buhrmester, Kwang, & Gosling, 2011; Chandler, Mueller, & Paolacci, 2014; Mason & Suri, 2011; Paolacci, Chandler, & Ipeirotis, 2010). Studies have shown that samples tend to be disproportionately young and female

https://doi.org/10.1017/iop.2015.23 Published online by Cambridge University Press

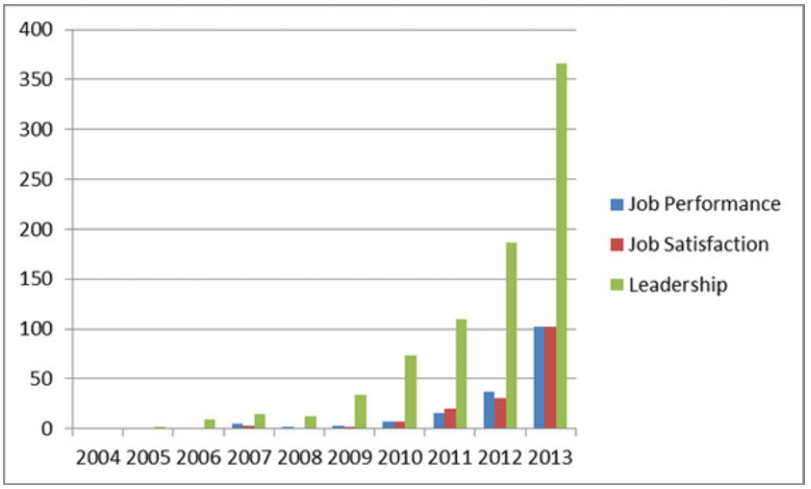


Figure 2 Google Scholar Search Results of Organizational Topics Within Studies Using Mechanical Turk (Color online)

with higher than average income levels (Buhrmester et al., 2011; Chandler et al., 2014; Paolacci et al., 2010).

The problems associated with representativeness become magnified when using non-U.S. samples. Not only do the percentages of individuals "working" on MTurk for a living dramatically increase but also the vast majority of the non-U.S. Turkers are male, under the age of 25, and much more educated than the average individual in their country (Ross, Irani, Silberman, Zaldiver, & Tomlinson, 2010). Recent research suggests that data provided from this latter group is particularly poor (Feitosa, Joseph, & Newman, 2015), and screening them out of studies using an MTurk country-of-origin screen can be problematic as there are a number of online sites dedicated to showing international Turkers ways to set up fake addresses and accounts in order to make themselves eligible for U.S.-only studies. Feitosa et al. (2015) estimated that 5%—10% of "U.S. participants" were actually from overseas. Similar results have been found in other studies (e.g., Berinsky, Huber, & Lenz, 2012; Rand, 2012).

In terms of what industries are represented, one recent study that we collected with over 1,000 U.S.-based Turkers found that the top three industries for Turkers were informational technology (16.1%), arts/entertainment (12.2%), and finance (12.2%). For comparison, the U.S. Census reports that these industries constitute 2.6%, 2.0%, and 5.1% of the workforce, respectively. Across all North American Industry Classification System codes, there was only a small ($r = .12$) correspondence between MTurk and the working population.

https://doi.org/10.1017/iop.2015.23 Published online by Cambridge University Press

It is also important that organizational researchers recognize that the majority of MTurk participants also tend to be underemployed or unemployed (Ross et al., 2010), and many report that MTurk is their primary source of income (Paolacci et al., 2010). This may be of particular concern to organizational researchers as many of these individuals report that they are "employed" part- or full-time, but they are either interpreting working on MTurk as being employed or are simply reporting being employed in order to participate in studies that pay well. That said, we should be clear that many Turkers do sincerely consider what they do to be a job like any other. Websites like *mturkgrind.com*, *wearedynamo.org*, and *turkopticon* have been used to advocate for MTurk worker rights. In particular, it has been suggested that the rates of pay for projects be benchmarked to the U.S. federal minimum wage. Researchers paying less than this rate may be contacted by self-appointed "labor reps" pointing them to these sites and threatening to contact university institutional review boards.

One consequence of having so many "professional" Turkers in the system is that a very small number of workers are responsible for a disproportionate amount of the activity on MTurk. One of the most comprehensive investigations of this problem found that the top 1% of the most prolific Turkers were responsible for 11% of the human intelligence tasks (HITs) and that the top 10% were responsible for 41% of the HITs completed on the service (Chandler et al., 2014). The website Deneme (2009) reported that the top 22% of Turkers were responsible for 80% of HITs. In other words, although Amazon advertises that there are over 100,000 active workers, in truth the majority of work is being done by a very small number of dedicated users. Moreover, because Turkers tend to specialize in particular types of tasks, this has enormous consequences for nonindependence of samples for researchers collecting questionnaire or experimental data. To illustrate, a recent set of six studies by Berinsky and colleagues (2012) found that 24% of the Turkers in the total sample appeared in at least two of the studies and 1% were present in five or more. Even more concerning is that the number of dedicated users is not equally distributed across samples. To investigate, we ran a study of 1,000 Turkers in five blocks of 200. To categorize workers, we used a database of eight studies consisting of over 20,000 Turkers that had been collected over the past 3 years. Individuals were coded as "unique" if they had taken part in no prior studies, "casual" if they had participated in one or two prior studies, and "dedicated" if they were in three or more. The results are presented in Table 1. As can be seen, the dedicated Turkers quickly inundate newly posted HITs. Additionally, 52.3% of our sample had participated in at least one prior MTurk study in our database, which is more than double Berinsky et al.'s estimate. It is only when studies are sufficiently large ($n > 800$) or have the advantage of a large database of

https://doi.org/10.1017/iop.2015.23 Published online by Cambridge University Press

Table 1 The Frequency of Dedicated Turkers in a Large Sample

User type	1st 200	2nd 200	3rd 200	4th 200	5th 200
Unique	58	80	107	107	123
Casual	102	89	77	77	65
Dedicated	40	31	16	16	11

prior studies to screen them out that one can be assured that the professional Turkers do not make up more than 10% of the sample. Individuals new to collecting data on MTurk can only gain this advantage by running several screening studies at great cost.

But is it really so bad to have large numbers of professional Turkers in a study meant to examine organizational phenomena? We believe that the answer is yes. As noted above, these individuals tend to have irregular employment status and are more likely to be from non-U.S. populations. Although these dedicated workers report making greater efforts to focus on HITs (Chandler et al., 2014), it is likely that they do so because they are aware of data quality checks and the need to maintain their reputations. Rather, Mason and Suri (2011) report that data quality actually falls among frequent participants. Similarly, the website Deneme (2009) reports that prolific Turkers tend to use a smaller range of response options than casual users do in order to complete surveys more quickly and maximize income. This results in greater peakedness in the distribution.

But high-frequency participants are not the only source of bad data in MTurk samples. Feitosa et al. (2015) found that data collected from non-English speakers was particularly problematic. Given the prevalence of both groups in smaller MTurk samples, this presents a real threat to the validity of data collected in this domain. Chandler et al. (2014) reported that only around 33% of research conducted using MTurk samples used data quality checks and that on average 15% of respondents were removed. However, the extent of the problem of bad data in MTurk samples is not really well known because most studies only conduct simple screens. For example, Rand (2012) reported that 4% of participants reported different genders across two studies. Feitosa and colleagues used two quality control items and ended up screening out approximately 9% of their sample. A more recent study used extreme scores on self-reported psychopathology scales to remove 10% of participants (Shapiro, Chandler, & Mueller, 2013). Others have reported having to reject 20%–50% of their participants for not taking sufficient care to respond accurately and honestly (Bohannon, 2011).

https://doi.org/10.1017/iop.2015.23 Published online by Cambridge University Press

Our own work on data cleaning methods (see DeSimone, Harms, & DeSimone, 2015) led us to conduct a more vigorous analysis of potential problems with MTurk data. One useful piece of information is that around 15% of potential participants log in to studies and self-select out before completing the survey. We assume that these are high-frequency Turkers checking on the difficulty of HITs before accepting the task. Although we have argued in this article that these individuals are less likely to provide high-quality responses, it is also true that any self-selection by participants in or out of a study is a potential threat to validity. Of the individuals who completed our survey, 4.9% responded to inventory items in less than 2 seconds on average, 9.5% responded to bogus items in such a way as to indicate that they were not reading them, and as many as 35% would have been screened out using standard cutoffs for the Mahalanobis D^2 screen that checks for patterns of aberrant responses. Based on this, it appears that prior estimates of the rates of poor-quality data in MTurk samples may have been underestimated.

Although more work needs to be done, it is clear to us that organizational researchers need to use caution before making use of MTurk or other online samples. At the very least, robust data cleaning procedures are needed. That said, we highly recommend subtle methods of conducting data quality checks (see DeSimone et al., 2015). Screening questions at the beginning of surveys are largely ineffective as Turkers have many public forums available to them where they report on not only the rates of pay and the amount of time needed to complete projects but also whether the researcher is using screens to check either eligibility or data quality (Chandler et al., 2014). We would also advise that the reasons for accepting or rejecting work be made very explicit before participants take part in HITs. We would even go so far as to say that investigators may want to consider paying for bad data as simply a cost of doing research online. Based on personal experience, it is not worth saving 30 cents to have someone call your house and threaten to drive from Colorado to Nebraska so they can stab you because you did not pay them.

On the whole then, we differ from Landers and Behrend (2015) in that we urge a cautious, skeptical approach to embracing new samples. In a sense, they are advising us to focus on avoiding making a Type 1 error. We should not reject unconventional samples out of hand because those samples are "neither better or worse . . . merely different." Instead, we would argue for moving forward more slowly so as to avoid the Type 2 error of falsely concluding that there is no difference between types of samples. This would seem to be the most prudent course to take considering that even Landers and Behrend have noted the paucity of evidence to support or disconfirm either conclusion.

https://doi.org/10.1017/iop.2015.23 Published online by Cambridge University Press

In conclusion, and in contrast to Landers and Behrend (2015), we believe that rather than solving the problem of the WEIRD (*Western, educated, industrialized, rich, and democratic*; Henrich, Heine, & Norenzayan, 2010) samples typically found in psychological studies, the samples provided by MTurk and other online sites have simply added another confounding variable in that many of the participants are spending a disproportionate amount of time (and their work lives) *online* (Ross et al., 2010). We do not yet fully understand what the consequences of using samples of WEIRDOs to investigate organizational topics (and psychology more broadly) will be. However, we do agree with Landers and Behrend (2015) that more attention must be paid to whether such samples are appropriate. In the case of MTurk, we may find that such samples are not of much use to organizational researchers outside studies of job-seekers, the recently terminated, the chronically unemployed, and disengaged employees misusing company resources to earn extra income. The real inconvenient truth is that samples really do differ both in terms of their quality and in the degree to which conclusions can be drawn from them. The time has come for organization researchers to begin seriously investigating whether the recent enthusiasm for crowdsourced data is warranted or desirable.

References

Berinsky, A., Huber, G., & Lenz, G. (2012). Using Mechanical Turk as a subject recruitment tool for experimental research. *Political Analysis*, *20*, 351–368.

Bohannon, J. (2011). Social science for pennies. *Science*, *334*, 307.

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*, 3–5.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaiveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavioral Research*, *46*, 112–130.

Deneme. (2009, December 21). Deneme: A blog of experiments on Amazon Mechanical Turk. http://groups.csail.mit.edu/uid/deneme/?p=523

DeSimone, J. A., Harms, P. D., & DeSimone, A. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*, 171–181.

Feitosa, J., Joseph, D., & Newman, D. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, *75*, 47–52.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–135.

Landers, R., & Behrend, T. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice*.

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavioral Research*, *44*, 1–23.

https://doi.org/10.1017/iop.2015.23 Published online by Cambridge University Press

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.
Rand, D. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179.
Ross, J., Irani, I., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in Amazon Mechanical Turk. In K. Edwards & T. Rodden (Eds.), *Proceedings of the ACM Conference on Human Factors in Computing Systems* (pp. 2863–2872). New York, NY: ACM.
Shapiro, D., Chandler, J., & Mueller, P. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*, 213–220.

Don't Throw the Baby Out With the Bathwater: Comparing Data Quality of Crowdsourcing, Online Panels, and Student Samples

Nicolas Roulin
University of Manitoba

In their focal article, Landers and Behrend (2015) propose to reevaluate the legitimacy of using the so-called convenience samples (e.g., crowdsourcing, online panels, and student samples) as compared with traditional organizational samples in industrial–organizational (I-O) psychology research. They suggest that such sampling strategies should not be judged as inappropriate per se but that decisions to accept or reject such samples must be empirically or theoretically justified. I concur with Landers and Behrend's call for a more nuanced view on convenience samples. More precisely, I suggest that we should not "throw the baby out with the bathwater" but rather carefully and empirically examine the advantages and risks associated with using each sampling strategy before classifying it as suitable or not.

In this commentary, I examine and compare original data obtained from the three types of convenience samples highlighted in the focal article: (a) crowdsourcing, (b) online panels with participants recruited by commercial panel providers, and (c) business students. I discuss differences with regards to sample composition (i.e., age, gender, ethnicity, education, and

Nicolas Roulin, Asper School of Business, Department of Business Administration, University of Manitoba, Winnipeg, Manitoba, Canada.

I want to thank Steve Binggeli, Joshua S. Bourdage, Vanessa Jacksch, Ute-Christine Klehe, Franciska Krings, and Rima Tarraf for their help in the data collection, as well as Franciska Krings for her feedback on a draft of the article.

Correspondence concerning this article should be sent to Nicolas Roulin, Asper School of Business, Department of Business Administration, University of Manitoba, 406 Drake Center, Winnipeg, Manitoba, Canada R3T 5V4. E-mail: nicolas.roulin@umanitoba.ca

https://doi.org/10.1017/iop.2015.23 Published online by Cambridge University Press