Journal of the Academy of Marketing Science (2018) 46:789–794
https://doi.org/10.1007/s11747-018-0587-4

EDITORIAL

"Keep on Turkin'"?

John Hulland[1] · Jeff Miller[2]

Published online: 9 May 2018

Over the past decade, there has been tremendous growth in academic research use of online samples, coupled with a shift away from traditional sample groups (especially undergraduate college students). In particular, researchers are turning to crowdsourcing platforms like Mechanical Turk (MTurk) to obtain convenient and inexpensive samples. Launched by Amazon in 2005, MTurk is "an online crowdsourcing labor market" (Levay et al. 2016). In their review of work published over four years in the *Journal of Consumer Research* (volumes 39–42), Goodman and Paolacci (2017) observed that 27% of all surveys and experiments in the journal were conducted using MTurk. Even more strikingly, this proportion of MTurk-based empirical work grew from 9% in the first of the four years they reviewed to 43% in the final year. Similar growth has been observed in other social science fields, including (Chandler and Paolacci 2017, p. 500) "more than 40% of papers published in the *Journal of Personality and Social Psychology* and more than 20%" in *Psychological Science*.

Use of MTurk in Marketing's top journals has varied both by journal and over time. As noted above, its use has grown dramatically in *JCR*. A systematic review of papers recently published in the *Journal of the Academy of Marketing Science* (*JAMS*), the *Journal of Marketing Research* (*JMR*), and the *Journal of Marketing* (*JM*) reveals that published work in *JMR* using MTurk has grown from 12% of all papers in 2014 and 2015 to nearly 26% in 2016 and 2017. Its use in *JM* (14%) and *JAMS* (8%) has been more limited, but it is also increasing.[1]

This explosion in the use of MTurk has fostered interest in understanding the nature of MTurk "workers" and assessing the extent to which the data they provide can be considered reliable and valid (e.g., Berinsky et al. 2012; Goodman et al. 2013). The results of this assessment are mixed, as discussed in detail in an excellent review by Goodman and Paolacci (2017). Perhaps most importantly, the issue of economic motivation for MTurk workers to misrepresent themselves has been found to be an important determinant of data quality (Wessling et al. 2017). Researchers can take steps to assess and limit this potential for misrepresentation, but it seems clear that MTurk is not a panacea for the sampling woes of researchers.

Interestingly, the use of MTurk samples is fairly non-existent in the commercial world, where online consumer panels are believed to yield superior data, but at a higher cost. The second author of this piece heads Burke, Inc., the 18th largest marketing research agency in the U.S. Burke does not use MTurk samples with any of its clients. Furthermore, he is on the Board of the *Insights Association*, providing him with the opportunity to interact with a large number of executives from many other top commercial marketing research firms, yet knows of none that uses MTurk for its research. There are likely a few corporate researchers that use MTurk, but it is not a commonly used option by the established marketing research community.

In the discussion below, after reviewing the strengths and weaknesses of MTurk we broaden the discussion of sample crowdsourcing to look at other crowdsourcing options, focusing in particular on one alternative (Google Surveys) that seems to provide more accurate responses (see McDonald et al. 2013). We then empirically demonstrate that use of GS can potentially yield a more representative sample than that obtained using either MTurk or an in-company convenience sample.

What we know about MTurk

The good

Use of crowdsourcing destinations such as MTurk to collect data provides a variety of advantages when compared to traditional

[1] One primary reason for these differences in MTurk use is rooted in the fact that both *JMR* and especially *JCR* publish more experimental-based research. Nonetheless, increased use of MTurk can be observed in all four journals over time.

✉ John Hulland
jhulland@uga.edu

[1] Terry College of Business, University of Georgia, C303 Benson Hall, Athens, GA 30602, USA

[2] Burke Inc., Cincinnati, OH, USA

samples (Goodman and Paolacci 2017). These include (1) reduced costs, (2) quick response, (3) greater participant diversity, (4) superior data quality, and (5) greater research flexibility.

Reduced costs The costs to recruit and compensate MTurk workers are much lower than associated with traditional research pools (e.g., Berinsky et al. 2012), and many administrative burdens (e.g., coordinating experimental sessions, subject travel) can be reduced or even eliminated. As a result, researchers are able to stretch limited budgets to complete more studies and/or use larger sample sizes (resulting in higher powered studies).

Rapid response MTurk studies can be quickly designed, launched, and administered. Thus, researchers can more easily complete multiple studies, with greater flexibility in designing meaningful follow-up enquiries and relevant probes.

Participant diversity When compared to student populations or other convenience groups, the MTurk workforce offers the promise of greater participant diversity. Moreover, this diversity can range across customer segments, industries, even cultures. As a result, with careful study design researchers can potentially increase both internal and external validity (Goodman and Paolacci 2017).

Data quality MTurk workers are compensated for proper completion of assigned tasks (including successfully passing screening and attention check questions) and are often rated by researchers following task completion. As a result, MTurk workers are generally strongly motivated to follow instructions and pay attention to study details. Furthermore, they tend to be more materialistic and to value money over time to a greater extent than other study participants (Goodman et al. 2013). Thus, MTurk workers are generally viewed as conscientious and agreeable, yielding reliable and psychometrically sound responses.

Flexibility In addition to the ability to accelerate data collection (as noted above), use of crowdsourcing groups like MTurk can facilitate a wider range of study designs than is typically possible in the lab setting. Although many researchers use MTurk to source experimental subjects, some scholars have suggested use of MTurk workers across a broader range of research designs. For example, Goodman and Paolacci (2017) suggest that use of MTurk can facilitate longitudinal studies, cross-cultural research, interactions between participants, and use of alternative measures.

The bad

In addition to the above advantages, researchers have noted some key potential disadvantages associated with MTurk use (Goodman and Paolacci 2017; Wessling et al. 2017). These include (1) lack of representativeness, (2) self-selection, (3) participant nonnaiveté, and (4) participant misrepresentation.

Non-representativeness Although crowdsourced groups are typically more diverse than convenience samples (e.g., students), they are not necessarily representative of the underlying population of interest, depending on the aims of the research. For example, Goodman et al. (2013) found MTurk workers to be lower on extroversion, emotional stability, and self-esteem than the general population.

Self-selection Differential payment rates across studies can influence MTurk worker participation, as can study novelty, study attractiveness, and study recency, all leading to potential self-selection biases. Use of screening questions to retain only those workers eligible for study inclusion (i.e., they fit the population profile) can also encourage participant misrepresentation, particularly in the case of studies offering higher rates of pay (see below for further discussion of this point).

Participant nonnaiveté Over time, active MTurk workers will complete many studies, becoming familiar with both classic research paradigms and well-established methodologies. These "professional" respondents are likely to differ from others in systematic ways. Furthermore, information exchanged on MTurk worker forums about specific studies may bias subsequent participants' responses. Research suggests that using nonnaive participants may reduce effects sizes (e.g., Chandler et al. 2015, but see Zwaan et al. 2018).

Participant misrepresentation Perhaps the biggest challenge facing researchers using MTurk participants is the potential for individual misrepresentation, defined by Sharpe Wessling et al. (2017), p. 211) as "when a respondent deceitfully claims an identity, ownership, or behavior in order to qualify and be paid for completing a survey or … study." For example, Downes-Le Guin et al. (2006) found that 14% of the participants in one of their studies claimed personal ownership of a Segway human transporter.

The incidence of misrepresentation can be substantial when the economic motive to participate is large. Wessling et al. (2017) describe three of their own research study experiences with MTurk misrepresentation, suggesting that it can be a real problem in some cases. For example, 17 % of the workers who characterized themselves as over-50 smokers to qualify as participants in a lung cancer treatment study also described themselves as active, under-35 athletes to qualify for a study of shoulder dislocation treatments!

The issue is not one of inherent dishonesty in online participants; in fact, Chandler and Paolacci (2017), p. 501) note that many studies show that MTurk workers "are no more dishonest than other people when completing experimental tasks." However, even when those people who are inclined to lie represent a small proportion of the overall population, they can become a large part of a study sample when the qualifications used to screen potential participants are overly focused. For example, if only 5% of the population are over-50 smokers (i.e., the study target), and 5% of the population is willing to fraudulently misrepresent itself to gain study access, roughly half of the final study sample will be frauds. In general, the rarer the incidence of a focal group in the population, the greater the proportion of study participants who are likely to be misrepresenting themselves (e.g., Jones et al. 2015; Miller 2006).

Some crowdsourcing alternatives

Various solutions have been proposed to deal with the MTurk-related concerns noted above (e.g., develop and manage an ongoing panel, pay a fair wage to all participants whether or not they pass screening questions), and are well-described by both Goodman and Paolacci (2017); Wessling et al. (2017). Other options include using existing consumer panels or relying on other crowdsourcing alternatives.

Although the costs associated with using existing panels from companies like Survey Sampling, Inc. (SSI) and Critical Mix are typically much higher than for MTurk (currently about ten times the cost, for a general population survey), researchers can expect a more representative sample to result. New companies such as TurkPrime and Prolific Academic are also offering researchers opportunities to survey screened panels at a reduced cost, although the representativeness of their samples are not yet well-established (Wessling et al. 2017).

One other crowdsourcing alternative that has received relatively little attention from academics is Google Surveys (GS). The difference between GS respondents and marketplace samples or mechanical samples is that GS respondents do not set out to take a survey, nor are they seeking a monetary reward. Instead, GS respondents wish to read an article and have to answer questions in exchange for accessing the content. As a result GS samples are not compromised by self-selection biases, the presence of professional survey-takers, or participant misrepresentation. GS determines the age, gender, and location of potential respondents based on their browsing history and IP address, and selects samples of respondents from a wide network of news, information, reference, and entertainment sites.

Although academic use of GS for research has been extremely limited, the sampling approach is well accepted commercially. Sostek and Slatkin (2017) report that an average of three million GS surveys are completed a month, with a typical survey containing five questions (GS permits the use of up to ten questions per survey), and 18% of the surveys include at least one screening question. Furthermore, GS response rates are high by today's standards, suggesting that respondents are very motivated to complete the surveys. McDonald et al. (2013) note that in conducted trials the average GS response rate was 23.1%, in contrast to industry response rates of less than 1% for most Internet intercept surveys, about 10% for telephone surveys, and 15% for Internet panels.

Given its constraints on the number of questions that can be asked, GS clearly offers less survey design flexibility than other survey sample sources. It is not really suited to conducting experiments, and is better suited to assessing incidence rates. GS's point estimates tend to be quite reliable and accurate. As McDonald et al. (2013) report, the accuracy of Google Surveys is better than both probability and non-probability based Internet panels "on three separate measures: average absolute error (distance from the benchmark), largest absolute error, and percent of responses within 3.5 percentage points of the benchmarks."[2]

An empirical example

In order to provide an illustration of the representativeness of samples obtained through different crowdsourcing approaches, we conducted a small empirical field study asking two incidence questions of individuals drawn from four distinct data sources: (1) Google Surveys (GS; $n = 101$); (2) Mechanical Turk (MTurk; $n = 101$); (3) a convenience sample of Burke employees (Internal; $n = 110$); and (4) Survey Sampling, Inc. (SSI; $n = 201$).[3] In the case of SSI, we attempted to have a "representative U.S. sample" answer our incidence questions. SSI controls for demographic variables such as age, gender, income, race, and geographic region, resulting in the higher sample size for this data source. Theoretically, results from

[2] All three authors (McDonald, Mohebbi, Sostek) work for Google. However, their conclusions are based on substantial, credible statistical comparisons that demonstrate good measurement accuracy.

[3] The sample sizes we employ in our field study are quite small. However, they are large enough to illustrate that significant differences in response incidence can be observed through the use of different data sources. The SSI sample size is larger in order to more accurately represent the underlying population.

Table 1 When did you last purchase a case for your mobile phone?

	GS (n = 101)	MTurk (n = 101)	Internal (*n* = 110)	SSI (n = 201)
In the past month	8%	8%	9%	14%
In the past 2–3 months	14%	23%	13%	14%
More than 3 months ago	49%	60%	58%	39%
I have not purchased this item	29%	9%	20%	33%

the SSI source should be closest to the "truth," whereas results using the Internal Burke sample should be the least accurate since the company's workforce is known to be more educated and affluent than consumers in general. Results using GS and MTurk should fall somewhere in the middle.

We do not mean to suggest here that the SSI sample will reveal the "true mean," but that it represents a presumably better option than both MTurk and the internal sample for deriving incidence estimates.[4] Online panels like SSI represent an expensive option for establishing incidence, but are also recognized as de facto standards in commercial research. SSI, like all non-probability samples, has its biases, but it is a widely employed commercial standard and therefore provides an appropriate benchmark for comparison.

The first incidence question we used was "When did you last purchase a case for your mobile phone?" Respondents could answer "in the past month," "in the past 2-3 months," "more than 3 months ago," or "I have not purchased this item." Results are shown, by data source, in Table 1. Overall, these responses vary significantly across the four sources (χ^2 (9) = 34.86, $p < .001$). We then compared the responses obtained from each of the other three data sources to those observed for the SSI sample. Both the MTurk and Internal sample responses differ significantly from those obtained using SSI (χ^2 (3) = 27.89, $p < .001$ and χ^2 (3) = 11.70, $p < .005$, respectively), whereas the GS and SSI responses only marginally differ (χ^2 (3) = 4.30, $p < .1$).

The second incidence question we used was "At which of the following stores did you shop for a mobile phone case?" Respondents could answer "AT&T (in-store or online)," "Verizon (in-store or online)," "Sprint (in-store or online)," "Apple (in-store or online)," "Wal-mart (in-store or online)," "Amazon.com," or "Other." Multiple responses were allowed. To simplify our analysis we combined the first three categories, with the results reported (by data source) in Table 2. Once again, the response differences vary significantly overall (χ^2 (9) = 87.72, $p < .001$). Furthermore, the SSI responses were significantly different from those obtained using the GS (χ^2 (3) = 19.59, $p < .001$), MTurk (χ^2 (3) = 48.13, $p < .001$), and Burke internal sample (χ^2 (3) = 32.29, $p < .001$). However, the GS results were less different from the SSI responses than either those obtained via the MTurk or Internal samples.

Table 2 At which of the following stores did you shop for a mobile phone case? (Select all that apply)

	GS (n = 101)	MTurk (n = 101)	Internal (n = 110)	SSI (n = 201)
AT&T, Verizon, Sprint	46%	16%	13%	66%
Apple	5%	7%	12%	7%
Wal-mart	11%	15%	6%	17%
Amazon	42%	78%	44%	31%
Other	12%	14%	12%	34%

At a basic level, these results suggest that in terms of responses to the specific questions we posed, MTurk does about as poorly as an internal company convenience sample, and that both do worse than GS. It is not entirely surprising that MTurk respondents are more likely to shop at Amazon, given that their incentive is paid within the Amazon network, but it does provide a stark example of the potential non-representativeness of the sample that can result when using that data source. The extent of non-representativeness of any convenience sample, including MTurk, is difficult to determine for any given survey or study, but our results illustrate that more carefully constructed samples from broader audience pools can minimize this particular data quality threat. However, cost and timing considerations may prevent the use of more rigorous sample options.[5]

Conclusion and recommendations

In the commercial research world, "fit for purpose" is a phrase that gets used a great deal with respect to sample source selection. Industry researchers are reluctant to conclude that any online sample source is either good

[4] Before GS became available, Burke used surveys of its own employees to obtain incidence estimates. It did this for cost reasons, despite recognizing the skewed nature of its workforce.

[5] Levay et al. (2016) provide evidence that it may be possible to statistically correct—at least in part—for MTurk sample biases. Specifically, they suggest the use of nine "broad, measureable features": age, race and ethnicity, gender, income, education, marital status, religion, ideology, and partisanship. In some cases this approach may provide a cost effective way to collect a representative sample while using MTurk.

Table 3 Recommendations for using MTurk and other crowdsourcing alternatives

Issue	Suggested practices[a,b]
Mismatch between study aims and achieved sample	• Choose appropriate crowdsourcing option ○ For general population experiments, MTurk is fine; for incidence estimates, Google Survey (GS) or online panels typically better. ○ MTurk sample biases can potentially be statistically corrected using nine broad features, if measured.
Handling respondent misrepresentation	• Pay a fair wage appropriate to the assigned tasks (i.e., neither too low nor too high). Pay all study participants; for those who fail to meet screening criteria for focal study, transfer to a different study. Run a short, inexpensive prescreen as a separate study to identify subjects for focal study. The required characteristics for the focal study need to be concealed in this presecreen. Consider developing an on-going, dedicated panel.
Managing MTurk workers	• Use TurkPrime and other information sources to identify MTurk workers who may have too much experience with task (i.e., they are nonnaive). • Monitor MTurk worker forums to uncover shared information relating to focal study.
Enhancing generalizability	• Within an overall project or paper, consider using respondents from different sources (e.g., MTurk workers for study 1, students for study 2). • Over-reliance on MTurk alone can limit generalizability.

[a] The practices listed here are examples, and not meant to be exhaustive

[b] Some of this material is adapted from suggestions in Chandler and Paolacci (2017); Goodman and Paolacci (2017); Levay et al. (2016); Wessling et al. (2017)

or bad, arguing instead that sample source selection should be matched with the specific needs of an individual study. A careful consideration of the relative strengths and weaknesses of alternative sample sources must be undertaken by the researcher to find the best fit. It is our contention that academic researchers need to adopt the same perspective. (This recommendation is summarized in Table 3, along with others culled from our review of recent papers discussing MTurk and other crowdsourcing options.)

For example, our results suggest that surveys about shopping behavior incidence rates should be placed neither with an Amazon audience nor with a convenience sample of relatively educated and affluent respondents (e.g., the Burke internal sample), whereas Google Surveys may prove adequate for providing reliable estimates of behavioral incidence. Yet use of MTurk may be completely suitable for studies regarding different types of attitudes or behaviors, or for research studying effect differences across experimental conditions. (Much of the existing work in Marketing making use of MTurk workers has been experimental.) Thus, while there is evidence that academic researchers can "Keep on Turkin'"[6] and in many cases feel confident with their results, use of crowdsourced samples needs to be guided by careful consideration of the appropriateness of the sample *source* for the specific research context under investigation.

[6] With apologies to Robert Crumb.

References

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*(3), 351–368.

Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are imposters. *Social Psychological and Personality Science, 8*(5), 500–508.

Chandler, J. J., Gabriele Paolacci, E., Peer, P. M., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science, 26*, 1131–1139.

Downes-Le Guin, T., J. Meehling, and R. Baker (2006), "Great Results from Ambiguous Sources: Cleaning Internet Panel Data," in *ESOMAR World Research Conference: Panel Research 2006*, Amsterdaam, The Netherlands: ESOMAR.

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research, 44*(1), 196–210.

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*(3), 213–224.

Jones, M. S., House, L. A., & Gao, Z. (2015). Respondent screening and revealed preference axioms: Testing quarantining methods for enhanced data quality in web panel surveys. *Public Opinion Quarterly, 79*, 687–709.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of mechanical Turk samples. *Sage Open, 6*(1), 1–17.

McDonald, Paul, Matt Mohebbi, and Brett Slatkin (2013), "Comparing Google Consumer Surveys to Existing Probability and Non-Probability Based Internet Surveys," White paper (https://www.google.com/insights/consumersurveys/static/consumer_surveys_whitepaper_v2.pdf), Google Inc.

Miller, Jeff (2006), "Research Reveals Alarming Incidence of 'Undesirable' Online Panelists," *Research Conference Report*, RFL Communications, In.c (Skokie, IL), September–October issue. (Retrieved from http://www.burke.com/Library/Articles/Jeff%20Miller%20RCR%20PDF.pdf).

Sostek, Katrina, and Brett Slatkin (2017), "How Google Surveys Works," White paper (http://services.google.com/fh/files/misc/google-surveys-whitepaper.pdf), Google Inc.

Wessling, S., Kathryn, J. H., & Netzer, O. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research, 44*(1), 211–230.

Zwaan, Rolf A., Diane Pecher, Gabriele Paolacci, Samantha Bouwmeester, Peter Verkoeijen, Katinka Dijkstra, and René Zeelenberg (2018), "Participant Nonnaiveté and the Reproducability of cognitive psychology," Psychonomic Bulletin & Review, in press.