MECHANICAL TURK: IS IT JUST ANOTHER CONVENIENCE SAMPLE?

R. Nicholas Gerlich, West Texas A&M University
Kristina Drumheller, West Texas A&M University
Robin Clark, West Texas A&M University
Meagan Brock Baskin, University of Tulsa

ABSTRACT

The study explores the use of an innovative technique for data collection – Amazon Mechanical Turk. Three studies utilizing the Theory of Planned Behavior to assess population behavior were used to compare behavioral outcomes between Amazon Mechanical Turk and general and specified populations. Results show that Amazon Mechanical Turk is a viable and generalizable sampling technique when a general population sample is needed. However, when specific populations are desired Amazon Mechanical Turk might be suboptimal.

KEYWORDS: Convenience sampling, Data collection, Amazon Mechanical Turk, Survey research

1. INTRODUCTION

Empirical research is undoubtedly one of the most important aspects of a social and behavioral scientist's career. We collect data from subjects on a variety of different topics to advance knowledge through publication. Although many of our management and psychology colleagues dream of large organizational samples of paired supervisor/employee dyads or entrepreneur interviews to help answer research questions, the fact is that the majority of social/behavioral scientists must rely on convenience samples to collect data. However, finding a good convenience sample can be difficult.

Convenience sampling is a sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher. There are several types of convenience samples, but most social and behavioral science researchers have utilized convenience samples of graduate and undergraduate populations. Despite the ease that researchers can and do use student samples, their prevalence in the literature is still much debated (Bello, Leung, Radenbaugh, Tung, & Witteloostuijn, 2009; Sackett & Larson, 1990). Specifically, the results of student-based samples have been questioned as to whether they can be applied to a general population of interest, for instance employee behavior in organizations (Bello et al., 2009). To combat issues of general worker population representativeness, many researchers have sought out alternative data sampling techniques, such as internet-based sampling. The purpose of this study will be to compare a relatively new and popular form of data collection called crowdsourcing. Specifically, using three studies which utilized the Theory of Planned Behavior, we will explore how well sample

populations from Amazon Mechanical Turk compared with other internet-based sample populations

1.1 Amazon Mechanical Turk

Amazon Mechanical Turk (MTurk) is an online crowdsourcing website that offers businesses and developers an innovative way to access an on-demand workforce. To date, there have been several studies investigating the design, use, and data quality of MTurk as a viable data sampling technique (Behrend, Sharek, Meade, & Wiebe, 2011; Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Casler, Bickel, & Hackett, 2013; Cheung, Burns, Sinclair & Sliter, 2017; Landers & Behrend, 2015; Mason & Suri, 2012; Mason & Watts, 2009; Paolacci, Chandler, & Ipeirotis, 2010; Smith, Roster, Golden & Albaum, 2016). Most of the current literature on MTurk has focused on appropriate uses of MTurk for data collection in the social sciences, as well as exploring the external validity of MTurk study results. Much like the concerns discussed previously about student convenience samples, researchers need to question whether the effects found in MTurk accurately represent the effects found in other sample populations (Berinsky et al., 2012).

To explore the external validity of MTurk samples, Berinsky et al. (2012) compared the demographic characteristics of collected samples of MTurk workers to the samples of previously published work. In addition, they attempted to replicate treatment effects of previously conducted experiments to assess sample generalizability. With regards to demographics, MTurk samples reported higher than general education than other adult samples. Additionally, the MTurk participants were found to be younger on average, as compared to the American National Election Studies (ANES). Several other studies have also explored the demographic make-up and found the sample characteristics are consistent and generalizable to the general adult population, with only minor differences (see Behrend, Sharek, Meade, & Wiebe, 2011; Buhrmester et al., 2011; Paolacci, et al., 2010; Ross, Iranim, Silberman, Zaldivar, & Tomlinson, 2010). Berinsky et al. (2012) further explored the external validity of MTurk samples on a variety of political and general attitudinal measures that could be used to compare these measures to the general population. Results of this study did not provide significance tests to assess actual differences; however, substantive differences between groups were found. Ultimately the authors concluded that while the samples do not perfectly reflect each other they are also not drastically different, and that MTurk samples are more representative than other convenience samples such as students.

Further evidence for the external validity of MTurk samples can be found in research conducted by Paolacci et al. (2010) who went beyond exploring demographic and attitudinal variables. The authors examined the differences between an MTurk sample, a traditional student convenience sample, and participants in online discussion boards on three classic judgment and decision-making experiments. Results found that MTurk samples did differ slightly in some cases; however, the differences in samples were consistent with result variability within the judgment and decision-making literature. Several other studies have also compared MTurk workers on a variety of menial and experimental tasks to assess the validity of worker behavior and found that MTurk workers exhibited similar accuracy and quality of work as other samples (Alonso & Mizzaro, 2009; Buhrmester, 2011; Snow, O'Connor, Jurafsky, & Ng, 2008). Despite the extensive research comparing MTurk workers to other samples on menial tasks and decision-making problems, the majority of behavior science research utilizes a variety of measures to assess individual/worker behaviors, attitudes, intentions, skills, and abilities. For research to utilize data

sources such as MTurk, it is crucial to assess whether MTurk sample results are representative to general population sample results when it comes to behavioral survey research.

2. MATERIAL AND METHODS

Three studies were conducted, with two samples each. In each study, two samples were solicited to complete an electronic Qualtrics survey: 1) an MTurk sample was solicited via the Amazon Mechanical Turk website and, 2) a sample was solicited via social media sites such as LinkedIn, Facebook, and Twitter, as well as other electronic media. In each individual study, the Qualtrics survey was identical for the two sample populations.

Each study used the theory of planned behavior (TPB) as the theoretical framework. TPB posits that individual intentions to act a particular way can be predicted by measuring behavioral attitude, perceived behavioral control, and subjective norms (Ajzen, 1985; 1991). Attitude is the positive or negative affect felt by the individual, perceived behavioral control is determined by the individual's perception of the ease or difficulty of engaging in the behavior, and subjective norms relates to others' attitudes. Essentially, the more favorable an individual's attitude and perception of social norms along with their perceived control over their behaviors, the more likely the individual is to act.

Extensions in the theory over time have led to important indicators more closely related to action including desire, intent and plan to behave a particular way (Shaw, Shiu, Hassan, Bekin, & Hogg, 2007). An individual is motivated to act through desire to act, expressed intentions to act, and planning to actually act in a particular manner. Similar attitudinal scales measuring motivations to shop or donate provided a strong foundation for comparing the MTurk and social media samples across the three studies.

2.1 Study Descriptions

The first study (the "Lowes Study") examined shoppers' planned avoidance of shopping at Lowe's home improvement stores following the chain's withdrawal of advertising on TLC's All American Muslim show. Data from the Lowes Study were collected in December 2011 and January 2012. The second study (the "Komen Study") measured participants' planned avoidance of donating to Susan G. Komen for the Cure following the organization's removal of financial support of Planned Parenthood. Data from the second study were collected in February 2012. The third study (the "Livestrong Study") focused on planned avoidance of donating to the Livestrong Foundation in the aftermath of Lance Armstrong's televised confession on Oprah. Data from the third study were collected in February 2013.

2.2 Participants

In the Lowes Study and the Komen Study, an MTurk sample and a general population sample were solicited. In the Livestrong Study, an MTurk sample and a purposively targeted sample toward individuals with a connection to the sport of cycling (as participant, sponsor, employee, coach, etc.) were recruited. In the Livestrong Study, survey invitations were posted in

sport-specific Facebook pages. Basic demographics of the six samples in the three studies (age, gender, and race) were collected, and appear in Tables 1-3.

2.3 Procedures

In each study, participants were directed to Qualtrics (either directly or via MTurk). Participants completed a demographics questionnaire containing items on the participant's age, gender, ethnicity, marital status, education, political affiliation, and state of residence. In addition, each study focused on a different controversial situation in which the authors were interested in the participant's planned behavior. Central to TPB is that individual intentions to engage in a particular behavior indicate the likelihood of actually engaging in that behavior in the future (Ajzen, 1985; 1991). TPB has been employed in a variety of disciplines and studies and thus provides a unique opportunity to explore the external validity of MTurk utilizing a commonly assessed behavioral theory.

In each study, a measure of planned behavior was administered across both samples. Following the procedure outlined by Shaw et al. (2007) and Arjen (2006), a survey was administered using 7-point bipolar scales to measure each component of theory of planned behavior: Attitude, Subjective Norms, Perceived Behavioral Control, Desire, Behavioral Intent, and Planned Behavior. Item wording in each of these measures was consistent with that of prior studies (e.g., Ajzen & Fishbein, 1980; Ajzen & Madden, 1986; Shaw et al., 2007). The six scales consist of two items each, except for Attitude (four items) and Perceived Behavioral Control (three items). Two of the Perceived Behavioral Control items were reverse-coded to maintain internal consistency. Summated scores were calculated for each of the six scales and Cronbach alphas were calculated for each sample in each study (see Tables 4-6).

The scale items measuring planned behavior used in each of the studies were adapted to fit the unique scenario addressed in the study. Wording of these items was as consistent as possible to that set forth in Shaw et al. (2007) and modified to reflect attitudes and intent toward avoiding donating (in the case of the Livestrong Study and Komen Study) or avoiding shopping (in the case of the Lowes Study). To ensure quality data in both MTurk and internet-based samples, the measures utilized included open-ended questions. Open-ended questions allow researchers to assess whether participants are taking care and consideration in their responses (Mason & Suri, 2012). In addition, the authors screened all submitted surveys for response time and response patterns.

3. RESULTS

Substantive differences between study group samples were assessed. Consistent with previous research, samples were comparable on demographic factors such as ethnicity. However, there were difference in both age and sex in some of the studies. In the Lowes Study there were substantively more females in the general sample than in the MTurk sample. Further, in the Livestrong Study, the cyclist sample was substantively older and consisted of more males than the MTurk sample.

T-tests for independent means were calculated within each of the three studies, using sample membership (MTurk or author-solicited) as the grouping variable to compare scores for the six summated variables. The results show that for the Lowes Study and the Komen Study only one variable was significant at the $p < .05$ level between the two samples. In the Lowes Study, the

Behavioral Intention variable was significantly different between the MTurk sample and the general sample. In the Komen Study, the Attitude variable was found to be significantly different between the two samples. A different pattern of results was found in the Livestrong Study. Interestingly, four of the six variables were found to be significantly different at $p < .05$ between the two samples (Attitude, Desire, Behavioral Intention, and Planned Behavior). The only variables that were not significant between the two samples in the Livestrong Study were Subjective Norms and Perceived Behavioral Control.

These results indicate that, in the case of the Lowes Study and the Komen Study, the use of MTurk as a sample recruitment tool was valid. In the case of the Livestrong Study, the results illustrate a not unexpected result: that when specific qualifications are needed among sample members (e.g., cycling affinity), MTurk is probably not a valid method of recruiting participants. However, when general population samples are in order, the results suggest that MTurk can provide the breadth of respondents that is desirable for a diverse sample.

4. DISCUSSION

The purpose of this analysis was to use three studies using TPB to explore the use of MTurk in behavioral survey research. Specifically, can MTurk samples provide representative results? Based on the finding from our studies, we can confidently say it depends! As a data source, MTurk provides a viable means for recruiting samples that closely match the general population (the Lowes Study and the Komen Study). However, when looking for specific sub-group opinions, such as the case in the Livestrong Study, MTurk does not provide an optimal pool of sub-group members. The fact that, in the Lowes Study and the Komen Study, there was one scale variable that was significantly different between the MTurk sample and the internet sample is not necessarily an indictment against MTurk samples, but rather an indication of potential limitations existing with any convenience sample.

The Lowes Study and the Komen Study are each very different from the Livestrong Study in that the latter included a very specific sample of those with cycling affinities (not a general population sample). The former involved samples with no stated connection to either the Lowe's home improvement stores or the Susan G. Komen organization. With regard to the Livestrong Study, the four significant differences reported are not surprising, especially since the cycling community in general felt betrayed by Lance Armstrong. The repercussions of Lance Armstrong's doping are thus likely to be felt even more profoundly among cyclists than non-cyclists. The implications are that an MTurk sample could potentially be useful for comparing a general populace sample to a purposive sample.

MTurk samples also offer great versatility in recruitment. In all our three studies, samples were limited to U.S. residents, although virtually any country could be included. More importantly, though, is that the MTurk samples in our studies were distributed across the U.S., with at least 48 states represented in each of the three samples. This claim could not be made of the author-recruited internet samples which were limited to social media invitations among friends and followers of the authors. Thus, regional bias can be minimized or perhaps even controlled for by using MTurk samples. In addition to sample diversity, MTurk offers great speed with which one can collect data. In both the Komen Study and the Livestrong Study, required sample sizes were reached in less than 24 hours. Collection in the Lowes Study took a little more time, but data collection was still complete within 5 days. The anonymity of MTurk samples is also desirable

vis-à-vis author-recruited samples; minimizing or eliminating bias because of association. This, in conjunction with geographic diversity, may help provide representative results.

While both the Lowes Study and the Komen Study showed one scale to be significantly different between the samples, the efficacy of using MTurk samples is not lost. The dependent variable in the TPB is the scale *planned behavior*; and in both studies, there was no significant difference between the samples with regard to their planned shopping or donating behaviors. It is also possible that the fundamental difference of donating to a charity vs. shopping at a retail chain may have contributed to the significant difference of the one scale in the two studies. Furthermore, it is possible that artifacts of the author-recruited internet sample in the Komen Study may have contributed to the significant difference reported between the attitude scale scores in the two samples: many participants recruited by the authors resided in the same state, a state known for its more conservative leaning and thus criticism of Planned Parenthood. Similarly, the significant difference in the behavioral intentions construct in the Lowes Study may be explained by the readily available retail options for shoppers. Whereas donating to any organization is strictly volitional, the acquisition of household items may be viewed as non-discretionary. The awareness of competing stores may have influenced these results, but at the same time, access to those competitors may have thwarted any intent to avoid Lowe's.

4.1 Limitations and Future Research

While this meta study extends the research regarding use of MTurk as a valid means of sample recruitment, it is not necessarily possible to draw final conclusions regarding the efficacy thereof. For example, it is possible that there could be a volunteer bias among MTurk workers who are simply willing to do anything in exchange for a modest payment. It is also possible that the author-recruited samples may be biased because participants are potentially friends or acquaintances. Still, the results and conclusions reported above appear to indicate that MTurk samples are not substantially different from general population samples, at least as it pertains to applications of TPB, and in fact may be better, given that the complete anonymity between researchers and participants.

This study is limited in that the three studies were cross-sectional. Longitudinal studies in each of the three cases might reveal differences between the sampling techniques. Future studies should focus on other comparisons of MTurk vs. general samples, in different contexts, and using different measures. Another area of concern that warrants future research is the mechanical nature of MTurk. Specifically, are participants human? While we made multiple attempts to screen participants for actual survey participation, it is possible that machine-based responders can compromise MTurk. Thus, future research could explore research ability to detect machine-based response versus human based responses.

5. CONCLUSIONS

In spite of the limitations above, the results reported appear to indicate the efficacy of MTurk as a means of sample recruitment. While the method is not without criticism, it is likely to be superior to relying on student-based samples in general, particularly for reaching organizational members and household consumers, as well as participants dispersed across a wide geographical area. The method also avoids any perceived pressure between researcher and students regarding participation, as well as artifacts pertaining to students being at a very different stage of life than

their slightly older adult peers. MTurk might also improve upon author-solicited samples that may reflect an implicit bias because of associations between the participants and the researchers.

REFERENCES

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl and J. Beckman (Eds.), *Action-control: From cognition to behavior* (pp. 11-39). Heidelberg: Springer.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179–211.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice Hall.

Ajzen, I., & Madden, T. J. (1986). Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. *Journal of Experimental Social Psychology, 22*(5), 453-474.

Ajzen, I. (2006, January). Constructing A Theory Of Planned Behavior Questionnaire. Retrieved from https://www.researchgate.net/profile/Icek_Ajzen/publication/235913732_Constructing_a_Theory_anned_Behavior_Questionnaire/links/56f00f4508aeae9f93e804b6/Constructing-a-Theory-of-Planned-Behavior-Questionnaire.pdf

Alonso, O., & Mizzaro, S. (2009). Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 workshop on the future of IR evaluation* (pp. 15–16).

Amazon Mechanical Turk (2013). *Company overview*. Retrieved from https://www.mturk.com/

Barger, P., Behrend, T., Sharek, D., & Sinar, E. (2011). I-O and the Crowd: Frequently Asked Questions About Using Mechanical Turk for Research. *The Industrial Organizational Psychologist*, 49(2), 11-17.

Behrend, T., Sharek, D., Meade, A., & Wiebe, E. (2011). The viability of crowdsourcing for survey research. *Behavioral Research Methods*, 43(3), 800-813

Bello, D., Leung, K., Radebaugh, R., Tung, R.L., & van Witteloostuijn, A. (2009). From the editors: Student samples in international business research. *Journal of International Business Studies, 40*, 361–364.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). *Using Mechanical Turk as a subject recruitment tool for experimental research*. *Political Analysis*, 20(3), 351-368. doi:10.1093/pan/mpr057.

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.

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, 2156-2160

Chueng, J., Burns, D., Sinclair, R., & Sliter, M. (2017). Amazon Mechanical Turk in Organizational Psychology: An Evaluation and Practical Recommendations. *Journal of Business and Psychology*, 32(4), 347-361

Hooghe, M., Stolle, D., Maheo, V., & Vissers, S. (2010). Why can't a student be more like an average person? Sampling and attrition effects in social science field and laboratory experiments. *Annals of the American Academy of Political and Social Sciences*, *628*(1), 85-96.

Landers, R., & Behrend, T. (2015). An Inconvenient Truth: Arbitrary Distinctions Between Organizational, Mechanical Turk and Other Convenience Samples. *Industrial and Organizational Psychology*, 8(2) 142-164

Mason, W. A., & Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavioral 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.

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

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in Mechanical Turk. *In Proceedings of the ACM conference on human factors in computing systems* (pp. 2863–2872).

Sackett, P. R., & Larson, J. R. Jr. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, *2nd ed.*, pp. *420-489*). *Palo Alto, CA*: *Consulting Psychologists Press*.

Shaw, D., Shiu, E., Hassan, L., Bekin, C., & Hogg, G. (2007). Intending to be ethical: An examination of consumer choice in sweatshop avoidance. *Advances in Consumer Research (ACR)*, Orlando, FL.

Smith, S., Roster, C., Golden, L., & Albaum G. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, 69, 3139-3148

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. *In Proceedings of the conference on empirical methods in natural language processing.*

Table 1.
Study 1 - Lowe's Demographics by percent

Trait	MTurk	General
Male	56.5	37.3
Female	43.5	62.7
Age	32.6	41.8
White	85.2	85.4
Black	3.6	2.5
Hispanic	2.7	5.7
Other	8.5	6.3

Table 2.
Study 2 - Susan G. Komen Demographics by percent

Trait	MTurk	General
Male	39.3	31.2
Female	60.7	68.8
Age	36.6	34.9
White	81.9	85.2
Black	4.3	2.8
Hispanic	3.3	7.4
Other	10.5	4.6

Table 3.
Study 3 - Livestrong Demographics by percent

Trait	MTurk	Cyclists
Male	61.4	78.5
Female	38.6	21.5
Age	32.25	46.5
White	79.8	92.9
Black	3.8	0.0
Hispanic	7.0	2.1
Other	9.4	5.0

Table 4.
Study 1 - Difference in Lowe's theory of planned behavior results across samples

Subscale	MTurk a	General α	t	*p*
Attitude	.89	.86	-.11	.92
Subjective Norms	.79	.86	-1.84	.07
Perceived Behavioral Control	.87	.75	.60	.55
Desire	.98	.97	1.36	.17
Behavioral Intent	.90	.87	2.17	.03
Planned Behavior	.94	.97	1.27	.21

Table 5.
Study 2 - Difference in Susan G. Komen theory of planned behavior results across samples

Subscale	MTurk α	General α	t	*P*
Attitude	.96	.95	1.93	.05
Subjective Norms	.56	.73	.29	.77
Perceived Behavioral Control	.76	.80	-.60	.55
Desire	.95	.92	.49	.62
Behavioral Intent	.86	.86	.99	.32
Planned Behavior	.94	.93	.37	.71

Table 6.
Study 3 - Difference in Livestrong Results theory of planned behavior results across samples

Subscale	MTurk α	Cyclists α	t	*p*
Attitude	.94	.98	-2.47	.01
Subjective Norms	.57	.61	1.87	.06
Perceived Behavioral Control	.79	.85	1.74	.08
Desire	.95	.95	-5.91	.00
Behavioral Intent	.91	.89	-5.48	.00
Planned Behavior	.93	.94	-5.78	.00