ARTICLES

Practice-based considerations for using multi-stage survey design to reach special populations on Amazon's Mechanical Turk

Victoria Anne Springer¹, Peter J. Martini², Samuel C. Lindsey¹, I. Stephanie Vezich³

¹ Adobe Systems, Inc., ² Heidelberg University, ³ University of California, Los Angeles and Adobe Systems, Inc.

Keywords: screening, hidden population, research, survey, mechanical turk, mturk

<https://doi.org/10.29115/SP-2016-0029>

Survey Practice

Vol. 9, Issue 5, 2016

The purpose of this article is to provide practice-based insight into the use of Amazon's Mechanical Turk (MTurk) online crowdsourced labor market to conduct survey research with targeted or "special" populations. We begin with an overview of MTurk, followed by a review of the use of MTurk in the social sciences, the multi-stage survey technique that has been implemented in our research, and the considerations that have informed our methodological choices. This article is based on the successful use of this approach in our work since January 2013 in both academic and applied research.

Introduction

The purpose of this article is to provide practice-based insight into the use of Amazon's Mechanical Turk (MTurk) online crowdsourced labor market to conduct survey research with targeted or special populations. We begin with an overview of MTurk, followed by a review of the use of MTurk in the social sciences, the multi-stage survey technique that has been implemented in our research, and the considerations that have informed our methodological choices. This article is based on the successful use of this approach in our work since January 2013 in both academic and applied research.

Background on Amazon's Mechanical Turk

Established in 2005, Amazon's Mechanical Turk (MTurk) is one of the most popular sources of online crowdsourced labor. Over the past 10 years, MTurk has also gained a reputation as a rich source for social science research subjects (Mason and Suri 2012; Paolacci, Chandler, and Ipeirotis 2010). As described by Martini and Springer (2014) the market aspect of MTurk is simple. "Workers" registered with Amazon search for, select and complete assignments posted by "requesters" (such as social scientists). These assignments are referred to as "human intelligence tasks" or "HITs." Workers are paid a requester-specified amount for each HIT that they complete, contingent on the requester's approval of their work. Thoughtful, accurate, and complete work is further incentivized through internal statistics that are automatically calculated for each worker. For example, the percentage of HITs completed, accepted or rejected work can be used to set inclusion or exclusion criteria for access to HITs (see Springer, Martini, and Richardson 2015, for more information on the use of qualifications). According to Amazon, the labor market that MTurk offers is based on the premise that there are still many things that human beings can do more efficiently than computers. That is, "workers"

can perform tasks that even the most advanced algorithms and sophisticated software cannot, including making nuanced decisions, expressing opinions, and thinking like a human being. For social scientists, this human aspect of MTurk is an invaluable asset.

MTurk as a Research Resource for the Social Sciences

According to Amazon, there are over 500,000 workers in the MTurk labor market, half of which are located in the United States. The majority of the remaining workers are located in India, but it is possible to access people from around the world – though they are present in far fewer numbers. This global access, along with the speed of data collection and low cost, has contributed to the growing reputation of MTurk as a source of research participants. Some informal evidence suggests that 16 of the top 30 U.S. universities collect behavioral data via MTurk (Goodman, Cryder, and Cheema 2013). Stemming from this growing popularity, the study of the platform itself has become a topic of interest to population researchers.

Overall, the tone emerging from the study of MTurk is an optimistic one. Research suggests that MTurk workers are similar demographically to other Internet-based research panels (Ipeirotis 2009, 2010; Ross et al. 2010) and are more representative of the general U.S. population compared to American college student participants and other Internet samples (Buhrmester, Kwang, and Gosling 2011). The comparison with American college students as research participants is a particularly meaningful contrast for experimental researchers who have historically heavily relied upon that readily available pool of subjects. In support of the quality of the results produced by MTurk-based samples, efforts to replicate classic studies have been successful using workers as research participants. This includes work in the fields of decision-making (Paolacci, Chandler, and Ipeirotis 2010), behavioral economics (Horton, Rand, and Zeckhauser 2011; Suri and Watts 2011), and political science (Berinsky, Huber, and Lenz 2012).

The outlook is also quite positive for population researchers. Attesting to the overall quality of the data produced by MTurk workers, studies have shown that the demographic information reported by MTurk workers is both reliable and accurate (Buhrmester, Kwang, and Gosling 2011; Rand 2012). Though these demographic traits are rarely a perfect match for U.S. estimates, they do not present a wildly distorted view (Berinsky, Huber, and Lenz 2012). In general, MTurk workers tend to be White, female, slightly younger, and more educated (Buhrmester, Kwang, and Gosling 2011; Berinsky, Huber, and Lenz 2012; Paolacci, Chandler, and Ipeirotis 2010). Taken together, this growing collection of methodological and empirical support seems to cast a positive outlook for the future of MTurk in the social sciences.

Representativeness and Access to Special Populations

In their extensive evaluation of the representativeness of MTurk samples compared to local convenience samples, Internet-based panel surveys, and elite national probability samples for political science research, Berinsky, Huber, and Lenz (2012) illustrated that the most promising and cautionary aspect of the MTurk sample comes from the same source: its diversity. Looking beyond the attractiveness of MTurk's convenience for reaching online research participants, scholars have begun to evaluate the utility of MTurk as a unique point of access for rare or hard-to-reach populations. Using online surveys for the study of hidden populations has been most often reported in the public health literature in the study of risky behaviors – with great success [see Duncan et al.'s (2003) study of illicit drug users]. Likewise, the use of Internet-based methods to survey marginalized populations did not introduce selection bias in a study involving gay and lesbian participants (Koch and Emrey 2001). On the contrary, the features of the sample of gay and lesbian participants matched national data on the characteristic of this group in the general population. Research continues to grow and attest to the advantageous use of technology to improve research with hidden populations (Shaghaghi, Bhopal, and Sheikh 2011).

Consistent with prior research on the overall representativeness of MTurk samples, Martini, Springer, and Richardson (in press) found that MTurk workers are not an inscrutable match for the demographic features of the U.S. However, their work supports the use of online approaches for accessing hard-to-reach or rare populations. These researchers found that MTurk was an excellent source of participants for these unique groups, including underrepresented sexual identities (i.e. lesbian, gay, bisexual, transgender and intersex people), minority religious groups, and rare health-related populations (i.e. intravenous drug users). Their research revealed that each of these groups was more common in the U.S. MTurk sample than in U.S.-based, nationally representative comparison surveys (World Values Survey, National Survey for Family Growth).

Applying Multi-Stage Survey Design to Special Population Research

In an effort to continue the advancement of the methodological rigor applied to MTurk-based research, Springer, Martini, and Richardson (2015) have developed a replicable multi-stage survey approach to reach specific populations or target groups using U.S. workers. The success of this approach (described below) was adapted in conjunction with Adobe researchers for identification and access to target markets for survey purposes. The practice-based recommendations that follow are drawn from the utilization of this approach in both academic and applied research.

Our process is as follows:

1. A HIT is posted that invites worker to participate in a general demographic screening survey (“We want to know more about you!”), used to identify members of the target group or special population. The language used is neutral in tone and provides only as much description as necessary to inform the worker of the type of task that will be required. The screening survey is typically structured to take around 5 minutes (a nominal amount of time), for which the workers are paid, regardless of their qualification for the second, target market-specific, survey. After completing the screening survey, all workers are provided with a completion code and instructions that direct them to return to the HIT to receive their compensation.
2. The second step of this two-stage process involves the following processes:
 1. Based on their screening survey responses, workers are identified who meet the qualifications to participate in the second survey, which is intended only for the target population of interest. This can be done using conditional rules, skip patterns, and other tools available through online survey software (e.g. SurveyMonkey or Qualtrics).
 2. Those who qualify are directed to an additional message indicating that they are eligible for another survey. They are then presented with the opportunity to participate in an in-depth secondary survey (typically ranging from 20 to 30 minutes), for which they receive additional, increased compensation. Admission to this secondary survey is based exclusively on whether or not specific selection criteria were met in the screening survey. Workers who choose to complete the second, in-depth survey are provided with further instructions for submitting their work for approval on MTurk. Additional information on various mechanics for this process using popular online survey software is available from the authors.

Case Study: Surveying Muslims in the United States

Employing this method, we obtained a sample of American Muslims for a survey about the negative effects of discrimination and stigmatization (see Springer, Martini, and Richardson 2015). Data collection occurred from June 2013 to January 2014 (7 months). During that time, 3,189 MTurk workers in the United States accessed the screening survey, 150 of which identified as Muslim (4.7 percent – compared to 1 percent expected in the U.S. population as a whole (Pew 2011). All 150 Muslims were given an opportunity to participate in a second in-depth survey; 116 (76.7 percent) did so. The sample obtained mirrored age and education trends that have been generally observed on MTurk. The sample was younger (62 percent under the age of 30) and more

educated (55 percent college educated or higher degree) than a national sample of Muslims obtained by Pew (2011): 36 percent under the age of 30 and 26 percent college educated or higher degree.

However, the denominational diversity of the sample paralleled national estimates. The majority of respondents were Sunni (60 percent), followed by Shi'a (13 percent), Sufi (6.2 percent), and Nation of Islam (4.4 percent). These numbers closely resembled the pattern observed by Pew (2011) in their report of Muslim Americans (65 percent Sunni, 11 percent Shia, and 15 percent indicated no specific affiliation). On this dimension, our approach produced a sample that approximated the same denominational diversity achieved by Pew using a standard random-digit dialing approach to reach survey participants. Taken together, our approach appears to have captured a denominationally represented group of Muslims and provided access to younger and more educated respondents that were not highly prevalent in Pew's sample. Future use of this approach may provide a strong complement to traditional survey techniques.

Utility of the Multi-Stage Survey Design

This multi-stage approach was developed to address several concerns.

Selection and Misrepresentation

The defining features of our target group are not explicitly stated in the language of the MTurk HIT. This has been done to diminish the likelihood of selection bias based on the use of enticing (or unenticing) language in the description of the task. By using neutral, but descriptive language, our intent is to prevent generating disproportionate interest in our studies by non-random groups of people. This approach has also been taken to curb the potential for intentional misrepresentation. That is, to prevent people from mimicking or falsely reporting the traits that we are looking for in order to be invited to participate in additional surveys.

Accordingly, we recommend that the title and description of the initial screening survey are informative, do not reveal the selection criteria, and are largely neutral in tone. For example, we would recommend against a HIT phrased as, "Fun research study for people who love to play sports after work!" This would likely draw increased interest from athletes – or those who are content to pretend to be for the amount of compensation offered in the HIT. A more neutral invitation to participate in a "Survey about your hobbies and leisure activities" provides some indication that there will be questions about things that they do in their free time but does not indicate what the target group will be. This method of screening is intended to minimize overt lying or misrepresentation to qualify for a higher paying survey. For our work with corporate research partners, we have also maintained an unbranded, generic presence as a requester in order to mitigate selection and misrepresentation concerns.

Homogeneity of Highly Restricted Samples

As was previously mentioned, MTurk compiles internal statistics on the performance of workers that complete HITs. This includes HIT approval rating, number of HITs completed, and a special status that encompasses an exceptional level of performance: Master worker. Sample restrictions (Qualifications) may be based on any of these aspects of worker performance or overall “Master” status. Careful consideration should be given to the decision to employ any exclusionary criteria through MTurk (Master status or otherwise). From a statistical perspective, employing systematic constraints may artificially decrease the variability present in a population. For tasks requiring high precision (e.g. image categorization) for which “Master” workers are selected, this is likely a desirable outcome. For social science researchers, the potential increase of homogeneity in restricting such samples may unintentionally impact results and fail to represent the attitudes, behaviors, or other outcomes of the worker population as a whole (compromising its generalizability to the general public).

Specificity of Screening Criteria and Rewarding Workers

To date, we have not utilized the “Qualifications Test” options available in the programming tools. Our reasons are twofold. First, the complexity of the process used to identify target groups or special populations of interest for the type of social scientific research that we have conducted has required the use of sophisticated survey software to construct highly specific selection rules. These rules are beyond the current capabilities of the “Qualification” tools provided in MTurk.

Our second reason is a largely philosophical one. We believe that an MTurk worker's time is valuable and that work undertaken in the service of our research should be paid. This includes being screened. From a research ethics standpoint, equitable treatment across workers who do or do not qualify for an in-depth survey also speaks to their right to receive compensation for the time they have already invested in our research (nominal though it may be). From a methodological standpoint, future work may examine other factors that may be influenced by the use of a required “Qualification Test” vs. our multi-stage screening process.

Conclusions

The practice-based insights provided in this work are drawn from the successful use of this multi-stage survey technique in over a dozen studies intended to reach well-defined target groups and special populations through Amazon's Mechanical Turk. There are a number of challenges inherent with this approach, many of which have been discussed. However, there is also great potential in the continued refinement of research approaches that utilize online crowdsourcing as a source of research participants. As the popularity of MTurk continues to grow, it is our hope that continued research on the

practicality, utility, and validity of using online crowdsourcing for social research will advance the further vetting approaches such as these and reinforce the rigor with which they are applied.

REFERENCES

- Berinsky, A.J., G.A. Huber, and G.S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.Com's Mechanical Turk." *Political Analysis* 20 (3): 351–68.
- Buhrmester, M., T. Kwang, and S.D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Cheap, yet High-Quality Data?" *Perspectives on Psychological Science* 6 (1): 3–5.
- Duncan, D.F., J.B. White, and T. Nicholson. 2003. "Using Internet-Based Surveys to Reach Hidden Populations: Case of Nonabusive Illicit Drug Users." *American Journal of Health Behavior* 27 (3): 208–18.
- Goodman, J.K., C.E. Cryder, and A. Cheema. 2013. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples." *Journal of Behavioral Decision Making* 26 (3): 213–24.
- Horton, J.J., D.G. Rand, and R.J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14 (3): 399–425.
- Ipeirotis, P.G. 2009. "Turker Demographics vs. Internet Demographics. A Computer Scientist in a Business School." 2009. <https://www.behind-the-enemy-lines.com/2009/03/turker-demographics-vs-internet.html>.
- . 2010. "Demographics of Mechanical Turk. CeDER Working Paper10-01," 2010. <http://hdl.handle.net/2451/29585>.
- Koch, N.S., and J.A. Emrey. 2001. "The Internet and Opinion Measurement: Surveying Marginalized Populations." *Social Science Quarterly* 82 (1): 131–38.
- Martini, P.J., and V.A. Springer. 2014. "Finding Hidden Populations through Online Crowdsourcing Methods." In *Poster Presented at the Annual Conference for the Society for Personality and Social Psychology*. Austin, TX, February 26–28.
- Martini, P.J., V.A. Springer, and J.T. Richardson. n.d. "In Press. Conducting Population Research Using Online Crowdsourcing: A Comparison of Amazon's Mechanical Turk to Representative National Surveys." *The Journal of Methods and Measurement in the Social Sciences*.
- Mason, W., and S. Suri. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavioral Research Methods* 44 (1): 1–23.
- Paolacci, G., J. Chandler, and P.G. Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5 (5): 411–19.
- Rand, D.G. 2012. "The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments." *Journal of Theoretical Biology* 299: 172–79.
- Ross, J., L. Irani, M.S. Silberman, A. Zaldivar, and B. Tomlinson. 2010. "Who Are the Crowdworkers?" In *Shifting Demographics in Amazon Mechanical Turk*. CHI EA_2863–2872.
- Shaghghi, A., R.S. Bhopal, and A. Sheikh. 2011. "Approaches to Recruiting 'Hard-to-Reach' Populations into Research: A Review of the Literature." *Health Promotion Perspectives* 1 (2): 86.
- Springer, V., P.J. Martini, and J.T. Richardson. 2015. "Online Crowdsourcing Methods for Identifying and Studying Minority Religious Groups." In *Digital Methodologies in the Sociology of Religion*, edited by S. Cheruvallil-Contractor and S. Shakkour. New York: Bloomsbury Academic.
- Suri, S., and D.J. Watts. 2011. "Cooperation and Contagion in Web-Based, Networked Public Goods Experiments." *PLoS One* 6 (3): e16836.