Zhang, Bingbing, and Sherice Gearhart. 2020. "Collecting Online Survey Data: A Comparison of Data Quality among a Commercial Panel & MTurk." *Survey Practice* 13 (1). https://doi.org/10.29115/SP-2020-0015.

ARTICLES

Collecting Online Survey Data: A Comparison of Data Quality among a Commercial Panel & MTurk

Bingbing Zhang[1], Sherice Gearhart[2]

[1] Donald P. Bellisario College of Communications, Pennsylvania State University, [2] Department of Public Relations, Texas Tech University (TX)

Keywords: amazon mturk, survey, panel companies, data quality

https://doi.org/10.29115/SP-2020-0015

Survey Practice

Vol. 13, Issue 1, 2020

Scholars seeking online data collection have sought the assistance of professional panels, which are commercial entities that recruit prospective research participants. Although not intended for recruiting participants, Amazon's Mechanical Turk (MTurk) service is a crowdsourcing platform that is being utilized for a lower cost. This study compares data collected from MTurk (N = 760) and a professional panel (N = 1,110) to assess aspects of data quality and respondent features. Results show that MTurk may produce better data quality based on completion rate and success in passing manipulation checks. Participants also differ in terms of educational attainment, age, and preferences of certain devices used for participation. Overall, results provide insight that researchers collecting online data should take into consideration before choosing a data collection platform.

Researchers collecting online data increasingly utilize professional panels (Antoun et al. 2016). Opt-in panels allow researchers to cost-effectively recruit respondents compared to offline recruitment (Antoun et al. 2016). Amazon's Mechanical Turk (MTurk) crowdsourcing platform provides an alternate avenue for participant recruitment in exchange for payment (Hitlin 2016). This study compares data collected from MTurk and a professional panel to assess data quality, respondent features, and devices used for responding. Results showcase practical implications for acquiring valid data when recruiting online.

Literature Review

Data Collection

Online participant recruitment has been adopted for their convenience, efficiency, and low cost (Paolacci, Chandler, and Ipeirotis 2010). Two popular means of online recruitment include professional panel companies and crowdsourcing websites, such as MTurk.

Panel companies. Commercial panel companies recruit individuals or entire households to voluntarily join a pool of prospective participants (Alreck and Settle 2004). They serve as a middleman between the researcher and participants to target participants based on client requirements. Panels collect personal information to identify qualified participants and reach out when members fit a target population, allowing respondents to earn payment (Craig et al. 2013). Panels check data quality using IP and email addresses to verify

identities, excluding respondents with missing data, or using researchers' feedback about respondents who provide invalid answers (Hays, Liu, and Kapteyn 2015).

Panels often provide nonprobability samples (Alreck and Settle 2004) and are commonly used for academic and market research (Smith et al. 2016). Demand for quick access to participants has produced this lucrative industry (Smith et al. 2016). While some panels recruit probability samples and assist when re-contact is necessary (Alreck and Settle 2004), convenience samples appear more common (Baker et al. 2010). Panels charge a fee for each response, typically about $4.00 (USD) per completion (Kees et al. 2017), which goes directly to the company that uses a portion of payment to compensate respondents. While expensive, panels aid access to participants from specific geographic regions or those with certain demographic characteristics (Roulin 2015). However, panelists who frequently take surveys might alter responses (Hillygus, Jackson, and Young 2014) and bring measurement error.

Crowdsourcing (MTurk). Crowdsourcing is done when large online groups contribute to a larger goal. In research, crowdsourcing is desirable because it can recruit a large number of participants (Litman, Robinson, and Rosenzweig 2015). Released in 2005, MTurk hosts a pool of online workers, known as MTurkers, who receive payment for completing human intelligence tasks (HITs) for monetary compensation through Amazon (Sheehan and Pittman 2016). MTurk is popular among researchers because of its convenience, rich subject pool, and low cost (Sheehan 2018). Completed HITs serve as an evaluation standard for workers' reputation and requesters can limit participants based on reputation score (Peer, Vosgerau, and Acquisti 2014). However, there is also concern about professional MTurkers who complete HITs for payment (Sheehan 2018).

Although different payment rates in MTurk may influence sample representativeness (Hulland and Miller 2018), MTurk has been found to provide quality data (Casler, Bickel, and Hackett 2013). MTurkers also have general computer knowledge (Kees et al. 2017) and pay more attention to surveys compared to panelists (Berinsky, Huber, and Lenz 2012). Enhanced attention may be because payment is not received until work is approved (Sheehan 2018) and the evaluation system may aid data quality (Peer et al. 2017). However, MTurkers are known to have faster completion times (Goodman, Cryder, and Cheema 2013) and have been found to not thoroughly read questions, producing data of low quality (Smith et al. 2016). This leaves suspicion that MTurkers may spend less time on experimental stimuli and perform poorly on manipulation checks.

Data Comparison

Nonprobability samples collected from panels and MTurk cause reliability and validity concerns, while also hindering the ability to generalize (Sheehan and Pittman 2016). Regarding demographics, MTurkers are known to be younger (Heen, Lieberman, and Miethe 2014) and population-based samples (Berinsky, Huber, and Lenz 2012). However, MTurk produces data that has higher quality compared to student samples and tend to be more diverse in income, education, and employment status (Berinsky, Huber, and Lenz 2012). Data quality indicators can include completion time, response rate, straight lining, and scale reliability (Tourangeau et al. 2018). The type of device used to complete data collection may be associated with quality. For example, participants spend more time completing surveys on smartphones (Revilla, Toninelli, and Ochoa 2016), and younger panelists utilize mobile devices for participation (Merle et al. 2015). Although limited panel/MTurk comparisons have investigated devices used for survey completion, more should be known about differences.

Research Questions and Hypotheses

To examine differences between professional panels and MTurk, the following research questions and hypotheses are posed:

> RQ1: Does completion rate differ between samples?
>
> RQ2: Does completion time differ between samples?
>
> H1: Participants from the panel company will review the stimuli for a longer period of time.
>
> H2: Participants from the panel company will have better performance in the manipulation check.
>
> RQ3: How do samples vary by demographics including: (1) sex; (2) education level; and (3) age?
>
> RQ4: Do samples use different devices to complete participation?

Methodology

Overview

Two web-based studies were administered in the United States during fall 2018. The surveys were identical in terms of procedures, length, and content and optimized for smartphone/tablet responses using Qualtrics. Data collection was intended to assess aspects of news consumption on Facebook. The only minor difference was a subtle change in questions inquiring about opinions on social issues, which were not part of this investigation.

Sample and procedure

Study 1. The sample was populated by panelists recruited by Research Now*,* now known as Dynata following a merger with SSI, a private company that provides samples based on client needs. Each participant was contacted to voluntarily participate between October 10 and 14, 2018. Upon completion, respondents were compensated through an internal point system. The price per respondent was quoted at $4.00. After negotiations, the cost for each completion was $2.50. A total of 1,374 individuals attempted participation, but 264 were excluded due to missing data. The total cost totaled $2,775 and took about four days to collect (N = 1,110). The larger sample size was collected to achieve the lower cost per participant.

Study 2. The MTurk sample was collected on August 2, 2018, and took less than three hours. Registered MTurkers could see the listed participation opportunity and opt to voluntarily participate if they met criteria (N = 760). Usable responses included those who fully completed survey participation. Upon completion, each respondent who successfully completed the HIT in the set time window and submitted their user identification code was paid $1.50. MTurk charged a fee of $0.05 to ensure respondents were Facebook users and a 20% service fee, totaling about $2.04 per respondent. While the intent was to collect 720 responses for the price of $1,548 through MTurk, 897 individuals accessed the survey. However, some respondents completed the survey but did not receive payment because participation was completed after the two-hour expiration period for the HIT or they failed to enter their user identification code. Regardless, all completed responses were kept for analysis.

Measures

Completion time. The completion time for each respondent was recorded in a number of seconds. This was converted to minutes, and an individualized score was assigned (M = 16.78, SD = 44.08).

Stimuli time. Qualtrics recorded the time spent on two stimuli, including a news story and a series of Facebook user comments. Time spent on the stimulus was converted to minutes to produce an individualized score (reading news article: M = 1.22, SD = 1.03; reading comments: M = .88, SD = .89).

Manipulation check. Participants were asked to report the news story topic viewed with a multiple-choice question. Responses were dummy coded (0 = incorrect, 1 = correct).

Demographics. Respondents were split between female (54.4%) and male (45.6%). Their average age was 42.14 years (SD = 15.73). Education was recorded by asking participants to indicate their highest level of education achieved (1 = *less than high school* to 7 = *doctoral/professional degree; M* = 4.06, SD = 1.46).

Device. Upon completion, respondents were asked about the device used. Options included (1) desktop, (2) laptop, (3) smartphone, (4) tablet, or (5) other. Two respondents chose "other" and were combined with "tablet" to form one category.

Results

RQ1 asked whether completion rate differed between samples. A completed response was considered one that went through the entire survey and fully completed demographic questions. Cross-tab analysis revealed a significant difference in completion rate (X^2 (1, N = 2,269) = 5.47, p = .02). Although 1,372 people initially participated through the panel company, only 80.90% (n = 1,110) of panel respondents provided completed responses. While 897 MTurkers initially began participation, only 84.73% (n = 760) completed participation. Hence, MTurk produced more completions than did the panel.

RQ2 asked whether completion time differed between samples. An independent samples t-test revealed MTurkers took fewer minutes (M = 14.60, SD = 46.10) to complete the survey than did panelists (M = 18.27, SD = 42.59), t(1544.40) = -1.74, p = .08; d = .25. However, the time difference was not statistically significant.

H1 predicted that panelists would review the stimuli for a longer period of time than MTurkers. A t-test compared time spent on stimuli and revealed that, although MTurkers spent slightly more time reading the news page stimulus (M = 1.25, SD = .72) than did panelists (M = 1.20, SD = 1.20), no significant differences were identified t(1837.31) = 1.06, p = .29; d = .05. Further, no significant differences were noticed in length of time spent on the second stimulus from panelists (M = .91, SD = 1.04) or MTurkers (M = .85, SD = .62), t(1834.61) = -1.56, p = .12; d = -.07. Therefore, H1 was not supported.

H2 predicted that participants recruited through the panel would perform better in manipulation checks than MTurkers. Cross-tab analysis revealed a significant difference in performance (χ^2 (1, N = 1,869) = 27.67, p < .001). Further review of the data revealed that 97.50% of MTurkers (n = 741) passed the manipulation check. Among panelists, only 91.60% (n = 1,017) passed the manipulation check. Therefore, respondents recruited through MTurk outperformed those recruited from the panel, and H2 was not supported.

RQ3 asked how MTurkers and panelists vary by demographics, including (1) sex, (2) education level, and (3) age. As seen in Table 1, cross-tab analysis revealed no significant difference in sex across platforms (χ^2 (1, N = 1,870) = 1.45, p = .23). RQ3b, which inquired about education level, was answered using a t-test. Data analysis revealed that panelists reported lower educational attainment (i.e., between some college and a 2-year college degree) (M = 3.86, SD = 1.54) than did MTurkers (M = 4.37, SD = 1.27) (i.e., between a 2-year and 4-year college degree) t(1804.59) = 7.82, p < .001; d = .36. RQ3c was

Table 1. Cross-tabs analysis of sex by recruitment platform

Device	MTurk	(%)	Research Now (Panel Company)	(%)
Male	359_a	(47.20)	493_b	(44.40)
Female	401_a	(52.80)	617_b	(55.60)

X^2 (1, N = 1,870) = 1.45, p = .23

Note: Means with different letter subscripts differed significantly at $p < .05$.

Table 2. Cross-tabs analysis of device use by recruitment platform

Device	MTurk	(%)	Research Now (Panel Company)	(%)
Desktop	309_a	(40.70)	135_b	(12.20)
Laptop	395_a	(52.00)	141_b	(12.70)
Smartphone	45_a	(5.90)	760_b	(68.50)
Tablet/other	11_a	(1.40)	74_b	(6.70)

X^2 (3, N = 1,870) = 834.02, $p < .001$

Note: Means with different letter subscripts differed significantly at $p < .05$.

answered with a *t*-test, which revealed that panelists were significantly older (M = 45.37, SD = 17.12) than were MTurkers (M = 37.41, SD = 11.99), $t(1867.04) = 11.82$, $p < .001$; d = -.52.

RQ4 asked whether MTurkers and panelists completed participation using different devices. Cross-tab analysis revealed the use of devices varied significantly (χ^2 (3, N = 1,870) = 834.02, $p < .001$). MTurkers were more likely to utilize a desktop (40.70%) or a laptop computer (52.00%) (see Table 2). On the other hand, panelists were more likely to complete participation using a smartphone (68.50%) or a tablet (6.70%).

Discussion

The purpose of the current study was to compare data collected from MTurk and a panel, which are both used for online participant recruitment and data collection. Findings indicate that, despite the higher price tag paid to the commercial recruitment industry, MTurk demonstrated better data quality. Participants from each platform were found to have several differences and results showcase important practical factors to consider when determining how to spend limited research funds.

Regarding completion rate, the higher cost per participant recruited through the panel did not reflect the enhanced cost. Although researchers pay per completion, the fee goes to the panel company who is responsible for compensating respondents (Kees et al. 2017). This is often done through an internal-reward system; the details of which are not shared with the client. Therefore, compensation may not be deemed worthy by respondents once they

have started participation, leading to quitting. On the other hand, MTurkers choose the task based on topic, time necessary for completion, and payment is sent to them directly.

While completion time did not differ significantly, both groups spent a substantial amount of time completing the survey. The large standard deviation indicates some respondents spent far longer completing the survey, which could be a by-product across platforms due to interest in the monetary benefit. For example, participants open the survey link with the intent to complete the survey at a later time. Some MTurkers in the current study claimed they completed participation but were not paid because they did not submit in the 2-hour window (Lee, personal communication, August 2, 2018). This raises concern about "professional survey takers" who are continually re-contacted with participation opportunities (Hillygus, Jackson, and Young 2014). It remains unclear whether this has a direct impact on data quality.

Participants recruited by both outlets were found to review the stimuli for nearly equal periods of time, demonstrating engagement. Despite criticism that data collection can be done quicker through MTurk than a panel (Goodman, Cryder, and Cheema 2013), faster data collection does not guarantee data quality. Yet, MTurkers were more successful in passing the manipulation check, indicating engaged attention. This supports claims that MTurkers devote more attention to completing tasks because payment requires approval (Sheehan 2018). MTurkers are aware of the international evaluation system, which may enhance workers' attention (Peer et al. 2017).

Samples were found to similarly feature an equal mix of male and female respondents. However, MTurkers had a higher level of educational attainment and lower age than panelists, aligning with existent findings (Heen, Lieberman, and Miethe 2014). MTurk has become a place for well-educated young people to earn income (Hitlin 2016). These differences indicate that researchers should look for trends in their areas of investigation to best target their desired population.

Finally, results showcased differences between the devices used. Specifically, MTurkers likely used a desktop or laptop while panelists relied on smartphones and tablets. Differences are important because mobile devices offer less control and screen size reduces functionality while increasing scrolling (Napoli and Obar 2014). These differences have implications for online data collection reliant on visual/auditory aids.

This study is limited by lack of understanding of completion time and its influence on data quality. While speedy completion may result in poor data quality, participants on different platforms have longer completion time. Also, completion time and manipulation checks might not be the only data quality indicators, which future research should further explore.

Submitted: January 09, 2020 EST, Accepted: October 13, 2020 EST

REFERENCES

Alreck, P. L., and R. B. Settle. 2004. *The Survey Research Handbook*. 2nd ed. Boston: McGraw-Hill/Irwin.

Antoun, Christopher, Chan Zhang, Frederick G. Conrad, and Michael F. Schober. 2016. "Comparisons of Online Recruitment Strategies for Convenience Samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk." *Field Methods* 28 (3): 231–46. https://doi.org/10.1177/1525822x15603149.

Baker, R., S. J. Blumberg, M. J. Brick, M. P. Couper, M. Courtright, and J. M. Dennis. 2010. "Research Synthesis: AAPOR Report on Online Panels. AAPOR Executive Council by a Task Force. AAPOR Standards Committee." *Public Opinion Quarterly* 74 (4): 711–81. https://doi.org/10.1093/poq/nfq048.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.Com's Mechanical Turk." *Political Analysis* 20 (3): 351–68. https://doi.org/10.1093/pan/mpr057.

Casler, Krista, Lydia Bickel, and Elizabeth Hackett. 2013. "Separate but Equal? A Comparison of Participants and Data Gathered via Amazon's MTurk, Social Media, and Face-to-Face Behavioral Testing." *Computers in Human Behavior* 29 (6): 2156–60. https://doi.org/10.1016/j.chb.2013.05.009.

Craig, Benjamin M., Ron D. Hays, A. Simon Pickard, David Cella, Dennis A. Revicki, and Bryce B. Reeve. 2013. "Comparison of US Panel Vendors for Online Surveys." *Journal of Medical Internet Research* 15 (November): 1–11. https://doi.org/10.2196/jmir.2903.

Goodman, Joseph K., Cynthia E. Cryder, and Amar 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. https://doi.org/10.1002/bdm.1753.

Hays, Ron D., Honghu Liu, and Arie Kapteyn. 2015. "Use of Internet Panels to Conduct Surveys." *Behavior Research Methods* 47 (3): 685–90. https://doi.org/10.3758/s13428-015-0617-9.

Heen, M.S., J.D. Lieberman, and T.D. Miethe. 2014. "A Comparison of Different Online Sampling Approaches for Generating National Samples." *Center for Crime & Justice Policy* 1: 1–8.

Hillygus, D.S., N. Jackson, and M. Young. 2014. "Professional Respondents in Non-Probability Online Panels." In *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas, 219–37. NJ: John Wiley & Sons, Ltd.

Hitlin, P. 2016. "Research in the Crowdsourcing Age, a Case Study." https://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/.

Hulland, John, and Jeff Miller. 2018. "'Keep on Turkin'?" *Journal of the Academy of Marketing Science* 46 (5): 789–94. https://doi.org/10.1007/s11747-018-0587-4.

Kees, Jeremy, Christopher Berry, Scot Burton, and Kim Bartel Sheehan. 2017. "An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk." *Journal of Advertising* 46 (1): 141–55. https://doi.org/10.1080/00913367.2016.1269304.

Litman, Leib, Jonathan Robinson, and Cheskie Rosenzweig. 2015. "The Relationship between Motivation, Monetary Compensation, and Data Quality among US- and India-Based Workers on Mechanical Turk." *Behavior Research Methods* 47 (2): 519–28. https://doi.org/10.3758/s13428-014-0483-x.

Merle, Patrick, Sherice Gearhart, Clay Craig, Matthew Vandyke, Mary Elizabeth Brooks, and Mehrnaz Rahimi. 2015. "Computers, Tablets, and Smart Phones: The Truth about Web-Based Surveys." *Survey Practice* 8 (6): 1–6. https://doi.org/10.29115/sp-2015-0028.

Napoli, Philip M., and Jonathan A. Obar. 2014. "The Emerging Mobile Internet Underclass: A Critique of Mobile Internet Access." *The Information Society* 30 (5): 323–34. https://doi.org/10.1080/01972243.2014.944726.

Paolacci, G., J. Chandler, and P.G. Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment & Decision Making* 5: 411–19.

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. "Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research." *Journal of Experimental Social Psychology* 70 (May): 153–63. https://doi.org/10.1016/j.jesp.2017.01.006.

Peer, Eyal, Joachim Vosgerau, and Alessandro Acquisti. 2014. "Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk." *Behavior Research Methods* 46 (4): 1023–31. https://doi.org/10.3758/s13428-013-0434-y.

Revilla, Melanie, Daniele Toninelli, and Carlos Ochoa. 2016. "PCs versus Smartphones in Answering Web Surveys: Does the Device Make a Difference?" *Survey Practice* 9 (4): 1–6. https://doi.org/10.29115/sp-2016-0021.

Roulin, Nicolas. 2015. "Don't Throw the Baby out with the Bathwater: Comparing Data Quality of Crowdsourcing, Online Panels, and Student Samples." *Industrial and Organizational Psychology* 8 (2): 190–96. https://doi.org/10.1017/iop.2015.24.

Sheehan, Kim Bartel. 2018. "Crowdsourcing Research: Data Collection with Amazon's Mechanical Turk." *Communication Monographs* 85 (1): 140–56. https://doi.org/10.1080/03637751.2017.1342043.

Sheehan, Kim Bartel, and M. Pittman. 2016. *Amazon MTurk for Academics: The HIT Handbook for Social Science Research*. Irvine, CA: Melvin & Leigh.

Smith, Scott M., Catherine A. Roster, Linda L. Golden, and Gerald S. Albaum. 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 (8): 3139–48. https://doi.org/10.1016/j.jbusres.2015.12.002.

Tourangeau, Roger, Hanyu Sun, Ting Yan, Aaron Maitland, Gonzalo Rivero, and Douglas Williams. 2018. "Web Surveys by Smartphones and Tablets: Effects on Data Quality." *Social Science Computer Review* 36 (5): 542–56. https://doi.org/10.1177/0894439317719438.