Crowdsourcing Quality Concerns: An Examination of Amazon's Mechanical Turk

Marc J. Dupuis
marcjd@uw.edu
University of Washington
Bothell, Washington, USA

Karen Renaud
University of Strathclyde
Glasgow, UK
Rhodes University, RSA
University of South Africa, RSA
karen.renaud@strath.ac.uk

Rosalind Searle
University of Glasgow
Glasgow, UK
rosalind.searle@glasgow.ac.uk

ABSTRACT

The use of crowdsourcing platforms, such as Amazon's Mechanical Turk (MTurk), have been an effective and frequent tool for researchers to gather data from participants for a study. It provides a fast, efficient, and cost-effective method for acquiring large amounts of data for a variety of research projects, such as surveys that may be conducted to assess the use of information technology or to better understand cybersecurity perceptions and behaviors. While the use of such crowdsourcing platforms has gained both popularity and acceptance over the past several years, quality concerns remain a significant issue for the researcher. This paper examines these issues.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**; • **Social and professional topics → User characteristics**; • **Applied computing → Law, social and behavioral sciences**.

KEYWORDS

Amazon's Mechanical Turk (MTurk), crowdsourcing, information technology research, quality control, human subjects research, qualitative data, quantitative data, surveys, open-ended questions

ACM Reference Format:
Marc J. Dupuis, Karen Renaud, and Rosalind Searle. 2022. Crowdsourcing Quality Concerns: An Examination of Amazon's Mechanical Turk. In *The 23rd Annual Conference on Information Technology Education (SIGITE '22), September 21–24, 2022, Chicago, IL, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3537674.3555783

1 INTRODUCTION

Online crowdsourcing platforms, such as Amazon's Mechanical Turk (MTurk), have been around since at least 2005 [11]. It allows researchers and others to place a task, referred to as a "human intelligence task" (HIT), on the platform and request a specified number of workers, referred to as "Turkers," to complete the work for a specified amount of compensation. These types of platforms help researchers easily overcome one of the more challenging obstacles in conducting human subjects research—the acquisition of a sufficient number of participants. Additionally, since Turkers reside in large numbers across the United States of America and elsewhere, it is simple to obtain a large number of participants quickly and efficiently for a cost that is often far less than other approaches [2, 13]. Nonetheless, quality concerns remain an important issue for the researcher. In this paper, we will explore some of these issues in more detail, including an examination of the experiences we have had in using this platform.

2 BACKGROUND

Conducting research that involves human participants poses many challenges, including seeking and obtaining human subjects approval, obtaining funding to provide fair compensation, acquiring quality results, finding enough participants to satisfy the requirements of the study with respect to sample size, and procuring a sample that is diverse on any number of measures (e.g., geographically). Crowdsourcing platforms address many of these challenges, while also introducing some of their own.

Prior to the advent of online crowdsourcing platforms, it was common for researchers to use student populations, such as sophomores in an introductory psychology class [12]. One discipline in which human subjects research is quite common, psychology, often requires their students to participate in a number of studies as part of their degree requirements. Compensation may or may not be provided. The participation of students in research studies does serve multiple purposes, such as the student obtaining first-hand experience what it is like to be a participant in a study and also providing a source from which researchers may obtain a sample. However, the level of homogeneity on a variety of demographics can be problematic. More often, these samples consist of undergraduate students attending a specific college in a specific location.

MTurk provides for a geographically diverse population with participants primarily residing in the United States of America. Similar platforms exist in other countries and continents, as well as the United States [13]. However, Turkers do tend to be more highly educated and are more likely to be White than the population at large [2]. Thus, this should be taken into account when seeking greater diversity on demographics related to education and/or ethnicity. Likewise, it is relatively easy to obtain a sample of several hundred or even over 1,000 in as few as 24 hours [2]. Compensation is determined by the researcher and presented to the potential Turker prior to them accepting the task. However, it is not uncommon for Turkers to be vastly underpaid for the work being asked of them [5]. Ethical considerations should be taken into

SIGITE '22, September 21–24, 2022, Chicago, IL, USA

ACM ISBN 978-1-4503-9391-1/22/09.
https://doi.org/10.1145/3537674.3555783

account with respect to providing fair compensation to individuals being asked to perform work. Some research has suggested that the median wage for Turkers may be as low as $2 per hour [4].

Generally speaking, quality has been considered to be quite good with MTurk workers and comparable to other sample types [8, 13]. However, it may not be appropriate to assess quality in the same manner (e.g., Cronbach's Alpha) that is done for samples obtained through other recruitment methods [13]. Our experiences support much of what is noted above with respect to MTurk, including how easy it is to obtain large sample sizes quickly, efficiently, and for an overall low cost. However, some types of studies may not lend themselves as well to MTurk and other crowdsourced platforms. Or, perhaps, these other types of studies may reveal more accurately some of the quality concerns that may be inadvertently overlooked and not easily detected through traditional quantitative statistical analysis techniques.

3 CASE STUDIES

We conducted two large-scale studies using the MTurk platform. These studies consisted of both quantitative and qualitative questions. Multiple quality control questions were inserted into each study as attention check questions. These were limited to the quantitative questions with the answer provided in the question itself. Logic was employed within the Qualtrics survey platform that ended the survey once an incorrect answer was provided to a quality control question.

This accomplishes three things: 1) It preserves the MTurk rating of Turkers since they will not be provided with the correct completion code and thus will not have their work rejected after submission (i.e., they will not submit their work for compensation in the first place); 2) As a result of the first item, researchers are less likely to obtain emails requesting a reversal on the rejection. Turkers can become quite persistent and argumentative when their work is rejected and it impacts their ability to obtain future work due to their lower overall quality rating, and 3) It helps make subsequent data analysis easier since it is clear which Turkers had their work rejected automatically. A problem with this approach is part of the intended outcome—Turkers not having their quality rating impacted. If their quality rating were to be impacted, it would perhaps discourage low quality work to a certain extent.

3.1 Study 1

In our first study [10], we set the quality threshold for who would be eligible to complete the task quite low. In order to be eligible, the Turker had to have previously completed at least 50 prior HITS with an approval rate of at least 95%. We obtained a sample of 1,072 after 73 of the responses were discarded through the use of logic due to failing one or more quality control questions. However, this is not the complete story.

We included two questions to determine if they should be asked additional questions related to their own personal experiences as it related to having experienced shame for cybersecurity incidents they were involved in within the context of an organizational setting. For the first question, 429 of the retained participants provided an answer that led them to being asked subsequent qualitative questions about their experience. However, there were many odd answers to these questions. It was not a matter of poor language ability, but rather answers that did not even begin to answer the questions being asked. This included several repeat answers to the questions. It appears participants were either using automated tools or manually placing the questions into search engines and then taking results from those searches and entering them in as their answers. Two raters were used to examine the results to the open-ended questions. If both raters agreed that it was more likely than not an illegitimate answer then the responses from that participant were discarded. Through careful textual analysis, we determined that out of the initial 429 participants answering this set of questions, only 53 (12.4%) provided usable responses to these open-ended questions.

For the second question, 342 participants provided answers that led them to the other set of questions. While less pronounced than what was found for the first question, a very small number (N=107; 31.3%) of usable responses were found. This number is incredibly low. And while much of this may be attributable to the relatively low Turker qualifications required to complete this HIT, the problem is not solved by simply increasing said qualifications.

3.2 Study 2

In our second study [9] (forthcoming), we opted to significantly increase the worker qualifications for those eligible to complete the HIT. Instead of 50 prior HITS at a 95% approval rate, we increased it to a minimum of 1,000 prior HITS with a 98% approval rate. There were 1,054 participants that began the survey with 1,000 successfully answering both attention check questions. As with Study 1, in Study 2 we had open-ended questions that were presented to the participants depending on their answer to a previous question. All participants were presented with some open-ended questions related to having experienced regret for cybersecurity incidents they were involved in, whether personally or within the context of an organizational setting.

Similar to Study 1, there were many responses that were not legitimate answers to the questions presented to them or they simply failed to answer the questions. Based on the same textual analysis approach noted in Study 1, we found 337 responses that had to be discarded. Thus, approximately 27% of the participants failed a quality control check either through the automated attention check questions or through the textual analysis, while another 9.9% failed to answer the open-ended questions or provided a simple one-word answer (e.g., good).

Overall, this marks a significant improvement from Study 1. Even if we were to give the complete benefit of the doubt to participants, 27% of them still failed quality control checks, which is significantly higher than the often touted 4% to 10% failure rate [2, 8, 13].

4 DISCUSSION

Quality control issues with MTurk are not new [1, 6, 7, 14]. There are online MTurk communities out there that provide a venue for Turkers to share HITS, provide feedback on requesters (e.g., researchers requesting the work), and identify the quality control question(s) for others. Virtual private networks (VPNs) are used to circumvent geographic restrictions on certain studies. Additionally,

various tools for automation may be found and employed by Turkers to increase the speed in which they can complete tasks.

Although quality control issues are a significant concern for researchers using crowdsourcing platforms, it does not mean that the use of such platforms should not be used. They provide a very important outlet for researchers to obtain data efficiently and cost-effectively—the value of which has not been matched by other means. Instead, it is important to be aware and develop as many different types of quality control checks as is reasonable. We did not use MTurk's "Master" workers for either study, which costs extra money with unclear benefits. However, it would be interesting to observe how they would perform in similar studies.

At the same time, it is important for researchers to provide fair compensation to Turkers. One way we assess this is by including a question in each of the studies for which MTurk is used and ask them how the compensation provided compared to similar projects on the MTurk platform: 1) Better compensation compared to others; 2) About the same, or 3) Less compared to others. Our goal that we have successfully met in each study is for 90% or more of the participants to indicate that the compensation received was about the same or better than similar projects.

Finally, it is important to note that in both studies we obtained a large number of incredibly high-quality responses from Turkers that genuinely wanted to provide thoughtful responses to our questions. The level of anonymity provided by MTurk for the Turkers also affords them with perhaps a greater level of comfort to share their experiences and perceptions. In face to face studies, even conducted virtually, this same level of comfort would not be present [3].

5 CONCLUSION

This paper examined quality control issues present in the MTurk platform. Additionally, it demonstrated how such issues may not be easily found in strictly quantitative studies. Thus, statistical analyses and conclusions therein are being drawn based on data that may have significant unknown quality control issues. After all, if this large number of Turkers are not answering open-ended questions in a thoughtful and/or truthful manner, it is likely safe to assume that the same is true for their responses to strictly quantitative questions.

Improving the requirements for Turkers to complete a HIT did help significantly, but it was not a panacea. Additional measures are needed. Other automated quality control measures may also be employed, such as asking them the same few questions (e.g., demographics) at both the beginning and end of the survey. If they do not match then it may suggest a quality control issue for that response. Crowdsourcing platforms are too valuable for us to throw the baby out with the bathwater; however, it is critical that we continue to evaluate their use.

REFERENCES

[1] Michael Chmielewski and Sarah C. Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science* 11, 4 (2020), 464–473.

[2] Marc Dupuis, Barbara Endicott-Popovsky, and Robert Crossler. 2013. An Analysis of the Use of Amazon's Mechanical Turk for Survey Research in the Cloud. In *International Conference on Cloud Security Management*. Seattle, Washington.

[3] Marc J. Dupuis and Karen Renaud. 2020. Conducting "In-Person" Research During a Pandemic. In *Proceedings of the 21st Annual Conference on Information Technology Education*. ACM, Virtual Event USA, 320–323. https://doi.org/10.1145/3368308.3415420

[4] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[5] John J. Horton, David G. Rand, and Richard J. Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* 14, 3 (2011), 399–425.

[6] Toni Kaplan, Susumu Saito, Kotaro Hara, and Jeffrey P. Bigham. 2018. Striving to earn more: a survey of work strategies and tool use among crowd workers. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[7] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas JG Winter. 2020. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods* 8, 4 (2020), 614–629.

[8] Gabriele Paolacci, Jesse Chandler, and Panagiotis Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making* 5, 5 (2010), 411–419.

[9] Karen Renaud, Marc Dupuis, and Rosalind Searle. 2022. Cybersecurity Regrets: I've had a few .... Je Ne Regrette. In *New Security Paradigms Workshop (NSPW '22)*. New Hampshire, USA.

[10] Karen Renaud, Rosalind Searle, and Marc Dupuis. 2021. Shame in Cyber Security: Effective Behavior Modification Tool or Counterproductive Foil?. In *New Security Paradigms Workshop*. ACM, Virtual Event USA, 70–87. https://doi.org/10.1145/3498891.3498896

[11] Oscar Schwartz. 2019. Untold history of AI: How Amazon's Mechanical Turkers got squeezed inside the machine. *IEEE Spectrum. Retrieved from https://spectrum.ieee. org/tech-talk/techhistory/dawn-of-electronics/untold-history-of-ai-mechanical-turk-revisited-tktkt* (2019).

[12] David O. Sears. 1986. College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology* 51, 3 (1986), 515.

[13] Zachary R. Steelman, Bryan I. Hammer, and Moez Limayem. 2014. Data Collection in the Digital Age: Innovative Alternatives to Student Samples. *MIS Quarterly* 38, 2 (2014), 355–378.

[14] Alex C. Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. 2019. The perpetual work life of crowdworkers: How tooling practices increase fragmentation in crowdwork. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–28.