Decentralized MTurk research during the 2020-2021 pandemic:

A means for students and emerging faculty to pilot novel studies, and learn research best practices

Patience Wieland and Aleshia Hayes

Department of Learning Technologies, College of Information, University of North Texas

Author Note

Patience Wieland https://orcid.org/0000-0001-7360-9755

Aleshia Hayes https://orcid.org/0000-0002-0880-077X

Correspondence concerning this article should be addressed to Patience Wieland, 16516 El Camino Real #424, Clear Lake City, Texas, 77062, United States. Email: patience@patiencewieland.com

Abstract

The 2020-2021 pandemic has created challenges for collecting data and recruiting participants, especially among emerging researchers. Amazon's Mechanical Turk (MTurk) platform is a low-cost solution for recruiting decentralized participants. The required learning curve for developing a pilot MTurk study, which includes developing an awareness of validation pitfalls, worker motivations and misrepresentation issues, has benefits beyond microworking platforms. Iterative design of a strong MTurk study also educates early career researchers. Lessons learned from MTurk can be applied to recruiting larger participant samples on social media and question and answer sites.

Keywords: Amazon Mechanical Turk, participant recruitment, research design, validation, coronavirus, COVID-19

Decentralized MTurk research during the 2020-2021 pandemic:

A means for students and emerging faculty to pilot novel studies, and learn research best practices

During the 2020-2021 coronavirus pandemic, we expanded the data collection for a previously designed study to include recruitment of research participants on the Amazon Mechanical Turk platform, an online platform that enables a variety of worldwide workers to complete business and research tasks for a fee (Follmer et al., 2017). When the study was first conceived, future participants were expected to come from a convenience sample, with participants recruited from two large universities in the United States.

Centers for Disease Control (CDC) and other public health guidelines have suggested that educational institutions use masks and social distancing to slow the spread of coronavirus disease (Honein et al., 2020; CDC National Center for Immunization and Respiratory Diseases (NCIRD) Division of Viral Diseases, 2021). Likewise, the CDC also recommended individuals avoid touching high-touch surfaces in public places. The public was advised to minimize handling cash, credit cards, and mobile or electronic devices when possible, which included virtual reality (VR) headsets.

Following public health guidelines also materially changed the kind of experiences we could explore in the previously designed study. This included our planned, repeated use of the same immersive technology devices by different study participants, a challenge particularly impacting immersive environment research (Steed et al., 2020). Using the Amazon Mechanical

Turk platform, commonly known as MTurk, enabled us to recruit participants from multiple countries that already owned personal VR devices and were available to join our study.

Literature Review

The Internet has produced a broader pool of potential research participants that can be recruited by graduate students and emerging faculty, not only through social media, but through the development of "microtask markets", such as Amazon's Mechanical Turk platform (Kittur et al., 2008). The system is named for an 19th century invention, a supposedly automated, chess playing machine that was controlled by an expert player, hidden inside (Bridges, 2014). Likewise, many of the tasks offered on this crowdsourcing platform could be completed by an artificial intelligence but benefit from the intelligence of humans; the tasks themselves are called HITs, for "Human Intelligence Tasks" (Wah, 2006).

A core benefit of the Mechanical Turk platform is the decentralized location of its worldwide users: early adopters such as Kittur et al. (2008) have noted the potential for generalizability across varied populations; Buhrmester et al., in a 2011 study exploring MTurk for psychological studies, were able to recruit over 3,000 participants from 50 countries and all 50 US states. Paolacci et al. (2010) note that Internet-recruited participants share greater similarity with the general American population, than student participants recruited on university campuses. Early explorations of validation suggested that MTurk participants would provide results roughly equivalent to convenience participants on American college campuses: while testing this hypothesis, Buhrmester et al. (2011) found that alpha levels ranged from .73

to .93 across differing compensation levels, and differing survey task lengths (from 5-30 minutes).

Conversely, we did not recruit on online gig platforms such as Fiverr, Upwork, and Freelancer.com. These platforms are venues where customers seek a match of skilled workers that can complete tasks and projects (Green, 2018). However, emerging professionals, and some experienced professionals with full time jobs, join these sites. Workers offer services or accept proposed gigs in the hopes of developing regular clientele, and to pursue intrinsically motivated goals such as owning their own businesses (Chen et al., 2019). Users who have joined these platforms to build communications-based businesses, such as managing social media or public relations tasks for clients, might be especially drawn to our VR project, which is about journalistic perception. These participants would be too knowledgeable of journalistic practices to provide a naive response to our experimental stimuli.

By contrast, the MTurk platform draws more workers who are primarily motivated by money, rather than self-improvement (Chen et al., 2019). This makes it an easier point of entry to recruit research participants who have at least a casual interest in new technologies like virtual reality, and technology-facilitated microwork, without expertise in the areas we are researching. While there are well over 500,000 MTurk workers registered (Follmer, 2017), only a few thousand are active; the half-life for individual user activity has been suggested as a year to 18 months (Difallah et al., 2018). This may be driven by the value workers initially find in being exposed to novel work experiences on gig platforms, and the correlation between greater job security and improved skills and reputation (Wood et al., 2019). As some MTurk users master certain kinds of online work, they may seek more specialized and regular business income on platforms such as Fiverr or Upwork, becoming less active on MTurk. Coupled with the younger age of MTurk workers (Chandler & Shapiro, 2016), these factors appear to improve our ability to collect a MTurk sample that is more representative of journalistic audiences, rather than expert practitioners.

A recent review of MTurk use by educational researchers (Follmer et al., 2018) suggests the platform is particularly suited for the study of college-aged and adult learners. These researchers also note that the popularity of educational research may also encourage sampling bias by MTurk users. Some users find they have implicit interests in specific areas of learning, such as metacognition and motivation, and may be unusually drawn to these kinds of studies. For this reason, Follmer et al. argue that "non-naïveté" should be considered for its impact on data and conclusions.

Wessling et al. (2017) describe experiences in which MTurk workers misrepresented themselves or attempted to "game" screening questions, to collect more desirable or higher paying assignments. Improved compensation has long been correlated with the ability to recruit quality participants on MTurk (Paolacci et al., 2010). Yet Wessling et al. (2017) draw attention to the increasing use of online forums such as Turkopticon and Hits Worth Turking For, which can mistakenly provide encouragement to overcome validity barriers, misrepresent oneself to capture valuable tasks, and discuss strategies to avoid being blocked from future tasks. Inattentive responses are another concern that have been raised, not only in collecting data from regular MTurk participants, some of whom may be multi-tasking, but also student samples collected online (Fleischer et al., 2015).

Results

During the coronavirus pandemic, we have been encouraged by the ability to recruit participants over the Amazon MTurk platform, and additional survey-friendly environments such as r/Samplesize, which is hosted on the social question and answer site Reddit. Our strategies tailored for collecting more quality participants on Amazon MTurk were easily adapted to other internet environments such as Reddit, suggesting that students and other emerging professionals may develop better practices through the learning curve required by MTurk studies. The extra preparation to ensure a more successful pilot study on MTurk with a small group of users, can encourage emerging researchers with rapid responses, but also enable them to test for validity challenges. This is particularly helpful when the researcher hopes to retool a high-stakes survey for launch across a larger number of sites, such as a dissertation study.

To successfully limit participants from the Amazon MTurk platform to high quality participants, including those who owned specific immersion devices, we redesigned the existing survey. First, we nested a CAPTCHA string in a pre-qualification survey, hosted on the Qualtrics survey tool. This pre-qualification strategy prevented automated submissions that could not get past CAPTCHA verification.

We also requested that MTurk users upload a unique picture of their personal immersion device. Using this strategy, we were able to remove potential recruits who had sent a variety of low-quality graphic images. We cross-checked additional device images against reverse image search, which enabled us to find scraped images of devices from pages on retail or journalism websites. This enabled us to limit further email recruitment to those who had sent unique pictures.

Instead of blocking users outright, which Wessling et al. (2017) have correlated with MTurk users engaging in character misrepresentation or other actions to maintain a lower task rejection rate, we used the MTurk qualification tool. In addition to demographic qualifications, this tool enables the work requester to attach one or more original labels to different MTurk users, which the workers themselves can see. For instance, a label called “VR Study I Qualification” could be attached to ten users that sent an image of their personal device and were recruited to the study. Another label, such as “Oak Tree,” or even a random string of letters and digits, could be appended to users who sent inappropriate, low-quality images. Importantly, because the users can themselves see this label, and because the MTurk worker audience includes a worldwide population that could misunderstand survey instructions, a researcher-created label should have what Amazon calls a “friendly name” (Amazon Mechanical Turk, 2017). A worker who is inappropriate for one survey may be useful in another. When the study is again listed as a potential task, both “VR Study I Qualification” users and “Oak Tree” label users can be filtered from seeing the assignment. By assigning a qualification to both those who pre-qualified for the study, and those who could not, we could prevent repeat attempts at survey recruitment tasks, improving validity. We could also, in a future study, use the tool to limit a follow-up survey or task to those who had previously been qualified.

Piloting these novel recruitment strategies on MTurk and Reddit, we were able to recruit 19 potential participants to our study in the first six weeks, all of whom sent photos of their personal devices. Zoom meetings have already been held with participants in five countries: Chile, Germany, Japan, Spain, and the United States.

A study that does not require ownership of specific, high-end technologies would likely be more successful in recruiting a variety of quality participants but can still benefit from the lessons learned here. By targeting higher quality responses on MTurk and responding to novel situations from our pilot recruitment efforts, we not only improved our recruitment numbers, but also the overall quality of our recruits. Our redesign for MTurk users improved the survey quality offered across other survey-friendly platforms. We successfully shared our revised recruitment messages on Facebook, Twitter, Discord, and several subreddits for virtual reality users. Tailoring recruitment for Mechanical Turk is an easy method for emerging researchers to rapidly improve their research design skills.

Significance

During the continuing 2020-2021 pandemic, temporary closings and public health guidelines for social distancing have made data collection more complex. We recommend that emerging researchers in teacher education carefully consider the Amazon MTurk platform and similar tools as a low cost means to recruit participants, and to pilot their design of surveys that will be shared widely over social media. While MTurk and other survey-friendly environments on social question and answer sites offer an opportunity for rapid recruitment of participants, careful planning and design are still crucial for study validity.

Design considerations should include the impact of non-naïveté (participants who are not naïve to the experimental stimuli) or topics that are unusually appealing to experienced users (Meyers et al., 2014). Additional considerations should include pre-qualifications, attention checks, reverse-coding scale items, incentive pay, recruitment materials and additional time budgeting to weed out low-attention submissions, and those who use automated means to attempt to capture higher quality or pay tasks. These strategies can ameliorate the value of the time committed to MTurk studies and the validity of participant responses.

References

Amazon Mechanical Turk (2017, September 17). *Tutorial: Understanding requirements and qualifications.* [Blog post.] https://blog.mturk.com/tutorial-understanding-requirements-and-qualifications-99a26069fba2

Bridges, E. (2014). Maria Theresa, "The Turk," and Habsburg nostalgia. *Journal of Austrian Studies, 47*(2), 17-36.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data*? Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 6*(1), 3-5. https://doi.org/10.1177/1745691610393980

Centers for Disease Control National Center for Immunization and Respiratory Diseases (NCIRD) Division of Viral Diseases. (2021). *Guidance for institutions of higher education (IHEs).* Centers for Disease Control. https://www.cdc.gov/coronavirus/2019-ncov/community/colleges-universities/considerations.html

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers*. Behavior Research Methods, 46*(1), 112-130.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12*, 53-81.

Chen, W., Suri, S., & Gray, M. L. (2019) More than money: Correlation among worker demographics, motivations and participation in online labor market. *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (Volume 13),* pp. 134-145.

Czerwinski, M., Lund, A., & Tan, D. (Eds.) (2008). *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*. ACM Press.

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of Mechanical Turk workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM 2018),* pp. 135-143.

Follmer, D. J., Sperling, R. A., & Suen, H. K. (2017). The role of MTurk in education research: Advantages, issues, and future directions. *Educational Researcher, 46*(6), 329-334. https://doi.org/10.3102/0013189X17725519

Green, D. D. (2018). Fueling the gig economy: A case study evaluation of Upwork.com*. Manag Econ Res J, 4*(2018), 3399.

Honein, M. A., Christie, A., Rose, D. A., Brooks, J. T., Meaney-Delman, D., Cohn, A., Sauber-Schatz, E. K., Walker, A., McDonald, L. C., Liburd, L. C., Hall, J. E., Fry, A. M., Hall, A. J., Gupta, N., Kuhnert, W. L., Yoon, P. W., Gundlapalli, A. V., Beach, M. J., & Walke, H. T. (2020). Summary of guidance for public health strategies to address high levels of community transmission of SARS-CoV-2 and related deaths, December 2020*. MMWR: Morbidity and Mortality Weekly Report, 69*(49), 1860-1867. https://doi.org/10.15585/mmwr.mm6949e2

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In M. Czerwinski, A. Lund, & D. Tan (Eds). *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08* (p. 453). ACM Press. https://doi.org/10.1145/1357054.1357127

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

Steed, A., Ortega, F. R., Williams, A. S., Kruijff, E., Stuerzlinger, W., Batmaz, A. U., Won, A. S., Rosenberg, E. S., Simeone, A. L., & Hayes, A. (2020). Evaluating immersive experiences during covid-19 and beyond. *Interactions, 27*(4), 62-67. https://doi.org/10.1145/3406098

Wah, C. (2006). *Crowdsourcing and its applications in computer vision.* Computer Vision Laboratory, University of California, San Diego.

Wessling, K. S., Huber, J., & Netzer, O. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research, 44*(1), 211-230. https://doi.org/10.1093/jcr/ucx053

Wood, A. J., Graham, M., Lehdonvirta, V., & Hjorth, I. (2019). Good gig, bad gig: autonomy and algorithmic control in the global gig economy. *Work, Employment and Society, 33*(1), 56-75.