

Using Crowdsourcing Websites for Sociological Research: The Case of Amazon Mechanical Turk

Daniel B. Shank

Published online: 14 March 2015

© Springer Science+Business Media New York 2015

Abstract Amazon Mechanical Turk, an online marketplace designed for crowdsourcing tasks to other people for compensation, is growing in popularity as a platform for gathering research data within the social sciences. Sociology, compared to some other social sciences, has not been as quick to adopt this form of data collection. Therefore, in this paper I overview the basics of Mechanical Turk research and suggest its pros and cons, both in general and in relation to different sociological data-collection methods and research needs. While Mechanical Turk is currently the most popular crowdsourcing website for research, I present general concepts, patterns, and suggestions that can be applied beyond Mechanical Turk to other crowdsourcing and online research.

Keywords Mturk · Online research · Methods · Data collection

Amazon Mechanical Turk (referred to as mturk; located at mturk.com) is an online marketplace designed to allow individuals or groups known as requesters to pay other individuals (workers) for completing small online tasks. In the last few years researchers in the social sciences, especially in psychology and economics, have utilized the mturk population as a way to collect research data. While this type of research collection has become more commonplace, sociologists do not frequently collect data using mturk or other crowdsourcing websites. First, I overview the basics of the mturk marketplace. Second, I answer the most prominent academic questions about using mturk workers for research: Who are the workers? Is the data reliable? What are the ethical issues? Then I suggest how sociologists can collect data using mturk, framed specifically for the diversity of methods used within sociological research.

D. B. Shank (✉)

Melbourne School of Psychological Sciences, Redmond Barry Building, The University of Melbourne, Parkville, VIC 3010, Australia
e-mail: daniel.shank@unimelb.edu.au

Amazon Mechanical Turk

Crowdsourcing, a combination of *crowd* and *outsourcing*, simply means to outsource a task or problem to other people. This can be as simple as asking a group of friends to each make one square of a quilt for a gift, emailing academic listservs to diversify one's literature search, or opening up the floor to an organization's members for suggestions. Using personal, professional, or paid networks, crowdsourcing is often desirable because it is quick and cheap, yet can lead to diverse, creative, or numerous results. Crowdsourcing is a fascinating social phenomena *to* research in that it is changing how people solve problems, network, innovate, and do business (see Shepherd 2012), but it also has potential uses *for* research with the advent of crowdsourcing marketplaces.

Online crowdsourcing marketplaces – those where people are paid to complete tasks – have only become popular in the last decade, and currently Amazon Mechanical Turk is the largest and most often used of them.¹ The name *Mechanical Turk* comes from the late 18th century Austrian Automaton Chess-player, a mechanical device that was dressed like a Turk. It was purported to be intelligent enough to play excellent chess; however its chess playing was in fact controlled by an unseen human. Amazon's Mechanical Turk, likewise, is designed for those with problems that require humans or groups of humans, not machines, to solve them.² Crowdsourcing problems to groups of people is often most effective when larger problems (e.g., transcribing an archive) can be divided into smaller tasks (e.g., transcribing a couple of pages) leading to the alternative names of microtasking or microwork for this type of problem-solving arrangement. Common tasks available on mturk include transcription, translation, photo and video tagging as well as marketing and academic research.

Here is how mturk works: Requesters have some task which they are willing to pay to have others complete. They set up their task on mturk.com either by using mturk's built-in tools, such as categorization, surveys, transcription from audio visual, or transcription from an image tasks, or by linking to another website which contains their task (Fig. 1). The mturk requesters include a description of the task, qualifications (if any) required of the workers, a time limit for completing the task, and the payment amount for the task. Then the requesters launch batches with a certain number of HITs (Human Interaction Tasks), the particular number of times they need that task completed. If they decide they need more or different versions, they launch additional batches.

After logging in, mturk workers can search and select among all requesters' HITs they desire and are qualified to do (Fig. 2). When they accept a HIT, they must complete and submit it within a certain amount of time or return it if they choose not to complete it. Requesters can accept (pay) or reject (not pay) for the submitted work, and additionally they can give bonus payments for excellent work. Requesters have names and workers have ids, meaning each can identify the other even without knowing their identity. If a worker does poorly, he or she can be rejected for a given task, or even blocked from completing future tasks by that requester. If a requester is

¹ Some others include ProlificAcademic, SocialSci, CrowdFlower, ClickWorker, CrowdSource. Some websites are specific to crowdsourcing particular types of tasks (e.g., graphic design) while others are general project markets like mturk.

² Mturk's motto is "Artificial Artificial Intelligence."

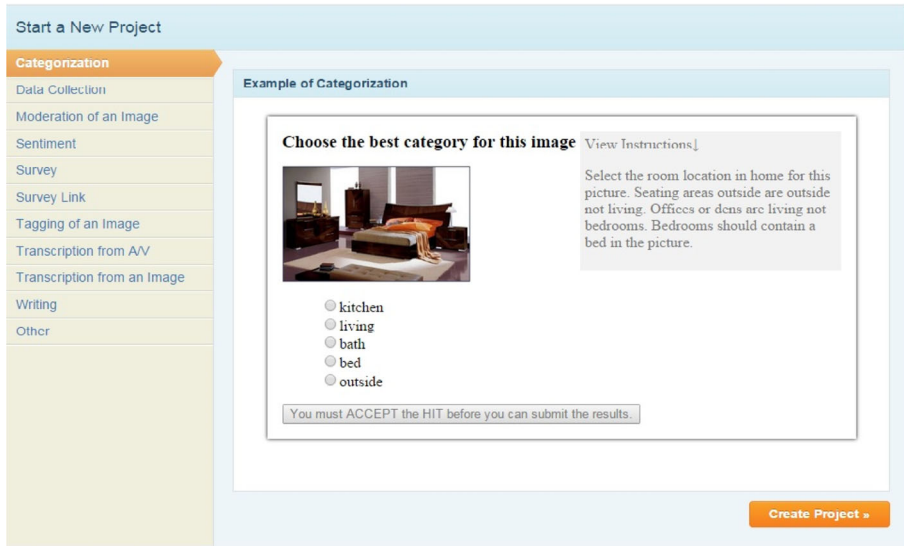


Fig. 1 Options for designing a tasks for workers

unfair or stingy, individual workers may stop completing his or her tasks. Both requesters and workers contribute to each other’s reputation through formal (e.g., task completion rates) and informal (e.g., blogs) means (see Mason and Suri 2012 for more details of mturk’s capabilities and process).

Tasks can require from a few minutes to several hours and compensation can range from a few cents to several dollars (with Amazon charging a 10 % fee to the requester). As a marketplace the pay rate varies considerably with some workers accepting tasks for a few cents, however at higher levels of payment requesters can potentially recruit hundreds of workers per day (Berinsky et al. 2012; Buhrmester et al. 2011). Finally, mturk contains built-in, and allows for custom-made, qualifications for tasks. A qualification allows the requester to specify what type of worker is needed for a particular task and can be based on the worker’s skills, characteristics, or levels of past performance.

Population of Mechanical Turk Workers

Mturk workers are from over 100 other countries, yet because mturk pays only in US dollars and Indian rupees (Mason and Suri 2012) the majority of the workers are from the US and India (Paolacci et al. 2010). The diversity of workers, especially US and Indian, allows for the possibility of cross-cultural comparative research (e.g., Eriksson and Simpson 2010). There are simple restrictions by country available when setting up



Fig. 2 Two HITs (Human Interaction Tasks) available to workers

mturk tasks, although any demographic characteristic can be prescreened by using mturk qualifications.

Studies of the US mturk population consistently find that mturk workers are younger, female, more liberal (Berinsky et al. 2012), and more educated (Paolacci et al. 2010) than the general US population. While the incomes of the US mturk population reflected the pattern of incomes in the general US population, they were generally lower (Mason and Suri 2012; Paolacci et al. 2010). Therefore, an mturk sample, like most samples, is not identical to the larger population, potentially due to individuals self-selecting into the mturk population based on financial incentives. Most participants, including those from other countries, do claim to be motivated primarily by money (Paolacci et al. 2010; Horton et al. 2011); however one study found that only 13.4 % of workers report mturk as a primary source of income (Paolacci et al. 2010).

Reliability of Mechanical Turk Workers

The most recent studies of mturk workers have found they produce reliable and quality data on well-established, mainly psychological, tasks. Buhrmester et al. (2011) compared the reliability of mturk workers on multiple classic personality measures and found they did not vary based on compensation levels. Furthermore, these data reached the reliability standards established for those measures for offline samples. Similarly, Horton et al. (2011) conducted framing, priming, and preference studies using mturk participants and found similar results to offline samples. Crump, McDonnell, and Gureckis (2013) replicated the results of a number of cognitive science tasks with millisecond-range reaction times. Paolacci, Chandler, and Ipeirotis (2010) compared an mturk sample with both offline and discussion board samples finding that the mturk sample did as well as, or better than, other samples in producing known social-psychological effects. Additionally an mturk sample paid higher levels of attention compared to other more traditional samples (Paolacci et al. 2010), established by a “catch trial” question which can identify those who are selecting answers without reading the questions (Oppenheimer et al. 2009).

With any sort of online research there is always some concern about participants’ true identities. While it is against mturk’s terms of service, individuals can potentially create multiple accounts, share accounts with others, or misrepresent themselves in other ways. Most scholars conclude that while there is always the possibility of misrepresentation online, it has been mitigated for mturk tasks by terms of service, qualifications, reputation mechanisms, and IP tracking (Horton et al. 2011; Rand 2012). Participants have been found to be fairly consistent both with their own answers across studies and also with self-report matched to external information such as IP addresses (Rand 2012).

Ethics and Ethics Approval for Mechanical Turk Research

Mturk research elicits the same types of ethical concerns as any other form of research. However, because crowdsourcing is a relatively new research platform, four issues warrant special attention, especially in applying to one’s Institutional Review Board or equivalent. First, informed consent and debriefing are similar to offline studies, with the information presented before and after the study within the mturk task. One can have

the participants read an online consent form and consent by checking a box in order to proceed. Second, workers are identified by their mturk ID, thereby making them initially anonymous. Therefore, anonymity may be easier to maintain in mturk studies compared to their offline counterparts, yet researchers must be careful about the ways in which information is revealed. For example, if an mturk worker emails the researcher then that email address and information associated with it would be known to the researcher.

Third, whereas anonymous identities are beneficial in some regards, they also mean that it is harder to identify members of restricted populations such as minors. I would suggest at minimum specifying in the consent form who is eligible to participate in the study. However Mason and Suri (2012) recommend an additional screening to prevent restricted populations from participating. Fourth, compensating is not restricted by minimum wage laws in the US as mturk workers are considered “independent contractors” (Mason and Suri 2012: 16). Different standards exist across disciplines, countries, and methodologies in terms of compensating research participants, so the ethically it is important to consider your own position as well as the mturk population which is primarily motivated by money, but not dependent on mturk for income (Paolacci et al. 2010; Horton et al. 2011). Some have suggested it is appropriate to compensate mturk workers at the same rate as the equivalent offline research (Crump et al. 2013; Mason and Suri 2012).

Sociology and Mechanical Turk Research

Does published sociology research include articles with data collected via mturk? I searched for articles mentioning mturk in a wide-range of generalist and specialized sociology journals³ and found only a few cases of mturk research. In total, 19 articles mentioned mturk, however on closer examination only 15 of these included data collected using mturk. Even among these 15 cases many were from more general social science or interdisciplinary journals that include sociology,⁴ therefore, it is clear that published sociology articles do not frequently include mturk data collection. Furthermore, there were no overviews in sociology journals of using mturk for research. In contrast, other disciplines have more rapidly embraced mturk to collect research data, usually using workers as participants in online experiments or surveys. Summaries, critiques, and investigations into using mturk for data collection are available in psychology (Buhrmester et al. 2011; Mason and Suri 2012; Crump et al. 2013), economics (Chandler and Kapelner 2013; Horton et al. 2011), political science (Berinsky et al. 2012), and interdisciplinary (Rand 2012; Paolacci et al. 2010) journals.

³ I searched in mid to late 2014 within any journals classified as (a) in the top 50 sociology journal by ISI Impact Factor, (b) an American Sociological Association official journal, (c) a regional US sociology association official journal, or (d) listed as a common sociology journal on several websites with such lists. Search criteria included using the words “mturk” or “mechanical turk.”

⁴ The fifteen research articles were in *Social Psychology Quarterly* (1 count), *Social Forces* (2), *Journal for the Scientific Study of Religion* (2), *Work and Occupations* (1), *International Journal of Intercultural Relations* (1), *Social Science Research* (4), and *American Behavioral Scientist* (4). Additionally, four articles in *Contexts* (1), *Politics and Society* (1), *Social Science Research* (1), and *American Behavioral Scientist* (1) mentioned mturk, but did not use data collected from it. For a comparison, psychology’s *Journal of Personality and Social Psychology* has 68 articles that at least mention mturk.

Sociologists employ a wide variety of methods in our research, and several of them were reflected in the 15 mturk data collection articles including a predominance of experiments or vignette experiments (e.g., Hunzaker 2014; Kuwabara and Sheldon 2012; Munsch et al. 2014; Simpson et al. 2013), two surveys (e.g., Johnson et al. 2012; Hart 2014) and a linguistic card sorting task (e.g., Ritter and Preston 2013). So while mturk research is underrepresented in sociology, the research does reflect methods used predominately by sociological social psychologists (e.g., experiments) compared to other sociologists. There are particular advantages and disadvantages in using mturk for different types of research, and so for the remainder of this section I will discuss the pros and cons of using mturk for different methods of data collection, moving from the most to the least common uses.

Experiments

Mturk has been widely used in other social science disciplines such as psychology and economics, and a few times in sociology, for experimental research. Because experiments focus on internal not external validity, experimentalists are primarily concerned with random assignment and control, not sample characteristics. An mturk sample being somewhat more like the general population, however, is often more diverse in age and socioeconomic status than a university sample (Berinsky et al. 2012; Paolacci et al. 2010). One advantage of experiments on mturk is that some experimenter bias and social desirability can be mitigated due to the lack of face-to-face contact (Paolacci et al. 2010). One disadvantage is lack of control over the environment as participants may be interrupted or distracted while completing the session. However, random assignment to experimental conditions typically means that any unforeseen effects from using mturk participants will affect all conditions equally. The exception to this is selective attrition, where participants in one condition drop out at higher rates due to condition differences; this can threaten internal validity for mturk experiments (Horton et al. 2011:415–416).

While many experimental paradigms make it clear one is participating in an artificially created task, others are more suited to be embedded within another task. Because mturk participants often do other types of tasks, those can be adapted into field experiment or serve as a cover story for an experiment (Horton et al. 2011). For example, an audio coding task could have participants count expression of emotions, interruptions, or the use of power in a conversation, with conditions varying the status characteristics of those in conversation. Chandler and Kapelner (2013) considered the differences in the framing of a task on participant's quality and persistence on that task giving either no context for a task, framing it as being important to medical research, or framing it as work that would be discarded. As they expected, participants were more attentive and produced higher-quality results when they believed the data would serve a valuable purpose.

Surveys

The mturk population varies from the general population on several dimensions as discussed previously. Therefore, mturk is not the best choice to obtain a nationally representative sample (see Berinsky et al. 2012 for a comparison of mturk surveys with

nationally representative and other surveys), although with some weighting or matching adjustments it is a possibility (Mason and Suri 2012). What is more promising for sociological research is the possibility of recruiting a specific, even hard-to-reach, population with an mturk survey. Mturk tasks have titles and keywords allowing one to target specific characteristics (e.g., survey for retirees) or interests (e.g., survey for those who play sports). It is also possible to use a short mturk qualification task to prescreening for some characteristics and invite participants that meet particular criteria to a second substantive survey. As mentioned previously, mturk has built in tools to design surveys, although one may also link the mturk participants to a survey hosted on another website (e.g., Qualtrics, SurveyMonkey).

To collect longitudinal data, specific participants may be invited to complete follow-up surveys by recording their mturk id or requesting their email address in the first survey. Those that routinely use mturk often enjoy having sustained interaction with the same requester, especially if they find the tasks interesting and the compensation fair. Therefore, the same participants could answer questions at different times, allowing for a reasonably cheap longitudinal dataset. Several scholars were able to get 60 % or better response rate for follow up surveys several weeks or even months after the original (Berinsky et al. 2012:note 9; Buhrmester et al. 2011). Yet I would caution scholars against longitudinal research that requires a timeframe of a year or more until more is known about the stability, repeat response rate, and non-response bias of this population.

Interviews, Focus Groups, and Observations

While survey questions on mturk can easily be quantitative or qualitative, this platform is not designed for collecting unstructured or semi-structured interviews, discussion or focus groups, or any type of naturalistic observation. No research that I am aware of employs any of these methods. It would be possible to have a chat-room style interview with someone who has agreed to participate, or to email follow up questions to a survey, but one would have to define these in terms of mturk tasks with payments as well as schedule times and procedures for direct online chatting.

Though mturk is not well-suited for collecting interviews, focus groups, and naturalist data, it is quite suited, and in fact designed, for coding them. Many mturk tasks involve coding or translating text transcripts, video, or audio, and qualifications available within mturk indicate workers who have exhibited skill and have experience in coding these types of data. Whether coding interaction behavior in a video, pauses or laughs in audio recordings, or simply transcribing recorded data to another format for easier analysis, mturk could be a cheap and efficient method.

A huge issue with such endeavors would be the privacy and ethical considerations of sharing these data with coders over the Internet. It would only be possible in situations where the data was not of a sensitive nature, and even then it would require a clear justification to receive clearance from an Institutional Review Board. Additionally, it would require the full approval of participants in the interview or interaction, and the disclosure that their data could be coded by hired others through Mechanical Turk. One option would be to allow participants to choose whether their data could be coded by a third party source such as mturk workers or would be restricted to coding by researchers and research assistants. Additionally, some naturalistic observations are

publicly available, so hiring mturk workers to code politicians' speeches or YouTube videos would be both efficient and offer no major ethical challenges. For a final note, other specialty websites specifically designed for coding of sensitive data (e.g., captricity.com) might be more suited than mturk for some projects, depending on factors such as costs, time-constraints, data sensitivity, the form of the data, and the size of the project.

Conclusion

Because research using Amazon Mechanical Turk provides access to a fast and cheap online population, there may be several situations where it is ideal. First, scholars who have limited research funding such as students, adjuncts, and some at teaching-oriented institutions may be able to collect data that is normally cost-prohibitive. Second, researchers who are planning large scale projects could benefit from a quick pilot study or seed research before submitting a grant or committing to a long-term data collection project. Third, academics concerned with specific populations might benefit from the ability to target particular characteristics of the online populations that may not be readily found in their local community.

For more information on becoming requesters, take mturk's tour at <https://requester.mturk.com/tour>. For those interested in more details about the process of doing social science mturk research, I would recommend Mason and Suri's article (2012) which specifies additional resources and details additional ethical, technological, and methodological concerns. In conclusion, I hope my brief overview will help sociologists to consider new opportunities for research using Mechanical Turk or other crowdsourcing websites. I believe that sociologists with our diverse approaches, methods, and substantive interests should be able to adopt this data collection tool to enhance the process of conducting research.

References

- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: amazon.com's mechanical Turk. *Political Analysis*, 20(3), 351–368. doi:10.1093/pan/mpr057.
- 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*, 6(1), 3–5.
- Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90, 123–133. doi:10.1016/j.jebo.2013.03.003.
- Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's mechanical Turk as a tool for experimental behavioral research. *PLoS One*, 8(3), e57410.
- Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, 5(3), 159–163.
- Hart, J. (2014). Did hurricane sandy influence the 2012 US presidential election? *Social Science Research*, 46, 1–8.
- Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14(3), 399–425.
- Hunzaker, M. B. F. (2014). Making sense of misfortune: cultural schemas, victim redefinition, and the perpetuation of stereotypes. *Social Psychology Quarterly*, 77(2), 166–184. doi:10.1177/0190272514521219.

- Johnson, M. K., Labouff, J. P., Rowatt, W. C., Patock-Peckham, J. A., & Carlisle, R. D. (2012). Facets of right-wing authoritarianism mediate the relationship between religious fundamentalism and attitudes toward Arabs and African Americans. *Journal for the Scientific Study of Religion*, 51(1), 128–142.
- Kuwabara, K., & Sheldon, O. (2012). Temporal dynamics of social exchange and the development of solidarity: “testing the waters” versus “taking a leap of faith”. *Social Forces*, 91(1), 253–273.
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon’s mechanical Turk. *Behavior Research Methods*, 44(1), 1–23.
- Munsch, C. L., Ridgeway, C. L., & Williams, J. C. (2014). Pluralistic ignorance and the flexibility bias: understanding and mitigating flextime and flexplace bias at work. *Work and Occupations*, 41(1), 40–62.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5), 411–419.
- Rand, D. G. (2012). The promise of mechanical Turk: how online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 299, 172–179. doi:10.1016/j.jtbi.2011.03.004.
- Ritter, R. S., & Preston, J. L. (2013). Representations of religious words: insights for religious priming research. *Journal for the Scientific Study of Religion*, 52(3), 494–507.
- Shepherd, H. (2012). Crowdsourcing. *Contexts*, 11(2), 10–11.
- Simpson, B., Harrell, A., & Willer, R. (2013). Hidden paths from morality to cooperation: moral judgments promote trust and trustworthiness. *Social Forces*, 91(4), 1529–1548.