LESSONS LEARNED

# RUNNING USER STUDIES WITH CROWD WORKERS

**Brian McInnis,** Cornell University
**Gilly Leshed,** Cornell University

Crowd work platforms are becoming popular among researchers in HCI and other fields for social, behavioral, and user experience studies. Platforms like Amazon Mechanical Turk (AMT) connect researchers, who set the studies up as tasks or jobs, to crowd workers recruited to complete the tasks for payment. Crowd workers on AMT (called Turkers) are quick and easy to recruit for online studies, are cheaper than paying people to come to the lab, and can provide useful feedback on prototypes through user research [1,2,3]. Plus, Turkers are considered more representative of the general (U.S.-based) population than the convenient undergraduate sample prevalent in academic research [4].

But behavioral and user studies on crowd work platforms can unearth challenges foreign to more traditional user research studies. For example, Turkers don't see themselves as study participants but rather as workers; they come to the AMT platform to do work and get paid, not to help with research. On the side of the crowd employer (called a Requester in AMT), it is easy to ignore Turkers; their work is mostly anonymous and the crowd work platform manages the labor arrangement and transactions, making it trivial to reject or even steal work [5]. A poorly designed experiment, such as a broken study platform or faulty survey questions, is difficult to detect because of the lack of direct contact between the researcher and participants. Further, such studies are often carried out quickly: A large number of workers can be recruited in a short amount of time, making it difficult to detect problems in the study until after many participants have engaged with it.

Here, we report on the lessons we learned about conducting research with crowd workers while running a behavioral experiment in AMT. We discovered the gray area of being both a researcher and an employer, and learned through trial and error what it takes to be a responsible researcher dealing with a large participant crowd. We hope that other researchers interested in using crowd platforms for user studies and behavioral experiments can learn from these lessons about treating crowd participants ethically and collaborating with them toward good results for the researcher and meaningful participation for the worker.

## SETTING UP A STUDY ON AMAZON MECHANICAL TURK

In 2014, we developed a research study to examine the behavior of first-time participants in online discussion forums and decided to use AMT as the platform for recruiting Turkers as participants. The study involved three stages: a pre-survey, a discussion forum, and a post-survey. Setting up a study in AMT uses a Human Intelligence Task (HIT)—the basic task that Turkers complete and for which they get paid. Researchers can set up simple studies using standard HIT templates, such as those for annotating photos or transcribing a recording. In our case, we needed to present Turkers with an interactive experience—the online discussion forum. We used a special type of HIT called an *ExternalQuestion*, presenting

IMAGE BY MOPIC / SHUTTERSTOCK.COM

the online discussion forum to Turkers through an iframe. An iframe is an HTML element that loads a foreign website as part of another. This enabled us to show the discussion forum and surveys within the AMT environment, without requiring Turkers to leave AMT and then come back to complete the HIT.

Before starting, we knew that Turkers were likely to share details about the task with others via online forums such as Turkopticon and TurkerNation. We also anticipated that some Turkers might try to cheat us and complete the tasks carelessly or maliciously—often referred to as "Spammers." What we did not expect was to find ourselves corresponding via email with several hundred Turkers who contacted us when encountering problems with the system. During the five months in which we prototyped the research system to run our study, we communicated with Turkers and found them helpful in setting up the study and the system, piloting it, debugging errors, and suggesting solutions.

## RUNNING THE STUDY AND HITTING CHALLENGES ON THE WAY

In the context of the study, we wanted to allow Turkers to interact with one another through an asynchronous discussion forum. Within the first few weeks of system development, we had a working discussion forum and had pilot-tested the study with small batches of three to 10 Turkers. These helped us improve the way we presented the study materials to the Turkers: for example, presenting comments entered by one Turker to others in the discussion forum, transitioning between stages of the study, and paying Turkers through the AMT protocols. We were comfortable with the system development and the study design, and felt ready to scale up the study.

On May 1 we launched the HIT, planning to recruit 90 Turkers to complete the online discussion study. Minutes after launching, we started receiving emails from Turkers describing errors they encountered in trying to complete the HIT. For example:

*I have completed your HIT on MTurk but finally unable to submit … Please find attached the screenshot of the problem I have faced. After I reached this stage, the button for continuing to the next page is not working. Please let me know what I should do.*

To our surprise, 137 started the HIT, and only four completed it. Being inexperienced with AMT, we learned that although we opened the HIT with 90 Turker assignments, not 137, when Turkers encountered errors and as a result abandoned the HIT, AMT recycled their assignments, making them available to other Turkers looking for HITs. This continued until we realized the system was broken and terminated the HIT.

By examining server errors and the emails we received from 33 Turkers, we were able to identify and fix some problems with the research system. We then continued to test different aspects of the study on AMT, each time with small numbers of Turkers.

On the morning of July 22, we launched a HIT to test the integration between the survey and the online

discussion forum through the iframe. The HIT initially seemed fine: From 6:16 a.m. to 6:51 a.m., everything was going smoothly. At 7:05 a.m., something went wrong. Ten Turkers emailed us in the next 20 minutes, reporting slightly different experiences with the survey. In total, 37 Turkers completed the HIT successfully, but some other Turkers were having a range of difficulties progressing through the surveys and the emails kept coming in. Again, we shut down the HIT. A few days later, after analyzing the experiences described in Turker emails and consulting with the third-party survey vendor, we found that the cause of these errors was rendering the survey inside of an iframe. The iframe was restricting the third-party survey from accessing the participant's browser, which was needed for the proper functioning of the survey.

## TECHNICAL ERRORS AND NEGATIVE FEELINGS

Beyond the technical errors in the research system, there were also personal difficulties in experiencing such errors, for Turkers and for us. To make the online discussion pertinent and interesting to Turker participants, we used the AMT Participation Agreement as the topic of discussion in the forum (available at www.mturk.com/mturk/conditionsofuse), prompting Turkers to discuss their experiences in relation to the Amazon policy. Turkers discussed issues such as HIT rejection, delayed payments, and Requester errors. When the system unintentionally broke, they experienced many of these issues firsthand. Specifically, the errors we introduced into the HIT prevented Turkers from completing the HIT and getting paid for their work.

Several Turkers initially thought that the error was deliberate or a joke, and they conveyed their anger in the emails they sent us: "Pretty upset because I spent probably close to a half hour on this HIT. It only pays $1 and I find it ironic since it was all about mTurk and fairness."

On Turkopticon—an activist technology designed to help Turkers identify good and bad Requesters by sharing their experiences—our Requester ratings plummeted, and we received comments that accused us of intentional wrongdoing: "How fitting that a HIT that talks about how Requesters can screw over Turkers is one that screws them over. Took 20 minutes to do and it didn't submit at the end."

Believing in transparency, we used our real full names, rather than pseudonyms, for our AMT account, which appeared near our HITs. As a result, the angry emails we received and negative comments on Turkopticon felt to us like personal attacks. The negative reviews drew the attention of Six Silberman, who together with Lilly Irani created and manage Turkopticon [6]. Silberman provided us with advice that was particularly meaningful and supportive: "Things can get really stressful in Turkland, with lots of people freaking out about accidents and assuming malicious intent at a moment's notice." A few Turkers we communicated with also sympathized with us: "If you have thick skin you can read up on the reactions of some of the people. This site [Turkopticon] is worker-controlled to fend off bad Requesters." Silberman reminded us not to take the angry comments personally, but instead to respond clearly and courteously, and to communicate persistently. As an honest Requester, it was important to us to resolve Turker concerns before continuing with more tests and eventually running the study.

## TURKERS DESERVE THEIR PAYMENT

Our first concern in recovering from the failures was to compensate the Turkers who started the HIT but because of our errors were not able to complete it. Similar to a lab study, participants expect fair compensation if they show up but cannot complete the study successfully because of the researcher's errors.

However, there is no way in AMT for a Requester to pay Turkers who started but did not finish and submit a HIT. Turkers who contacted us recommended a workaround for treating Turkers fairly in this situation: a "dummy HIT." The workaround uses one of AMT's standard template HITs, listing all the WorkerIDs associated with the Turkers we allow to accept the HIT. Once the Turkers completed this dummy HIT, we were able to pay them through a bonus.

We publicized the dummy HIT by posting on worker forums, emailing back those who emailed us when the original HIT broke and asking Turkers to share it with other Turkers. Unfortunately, we were unable to reach many Turkers who experienced the error and did not complete the HIT, because the AMT API does not include a way for Requesters to find and communicate with Turkers who abandon a HIT.

All of the Turkers who emailed or who had worked with us in the past were compensated for their participation in the broken HIT, and many emailed back supportive and encouraging comments, some being surprised to hear back from us after the error: "This is very much appreciated. Thanks for putting in the effort." We saw similar comments on Turkopticon: "Stand-up guy to work on it for so long to make sure people were compensated!"

## RECOVERY WITH THE HELP OF TURKERS

Our second concern was to fix the research system to be able to run the study successfully. When launching a HIT to run a study session, we learned to pay close attention to the server and database performance and to our email inbox, and to shut down the HIT early before errors get out of control. In response to Turker emails, we developed a routine of writing back a short standard message stating that we were working on the issue, then later, when we had more time, write back personally to each Turker. Both the initial stock reply and the later personal messages, although time-consuming on our end, were important for maintaining good relationships with Turkers; the messages helped us build rapport with them and proved useful as they helped us figure out some errors and how to recover from them.

In their initial emails when encountering a broken HIT, Turkers often provided us with screenshots, error messages from their browser consoles, and additional information about their experiences of the system. These details were a way for them to tell us that the errors were *not their fault*, and also helped us trace the source of the errors in order to fix them. We thought of these Turkers as our bug testers or consultants; they became valued collaborators: We contacted them ahead of some pilot tests, compensating them with additional bonuses as their work warranted. We continued to correspond via email with Turkers about the interface design and study instructions, and as

we incorporated major changes to the research system.

Based on advice from Silberman, we also decided to consult directly with veteran Turkers through the TurkerNation Internet-relay chat. Upon entering the #turkernation IRC, a few Turkers exited the space, stating that the "Big Bad Requester" had scared them off. However, others at the IRC were very helpful and gave us ideas about how to fix the broken system. Turkers suggested that instead of moving participants through the study within the iframe, we could provide a link to outside the AMT space to complete the survey. They encouraged us to trust Turkers to leave the AMT environment and then come back with a code generated by the survey to prove they had completed it. This extra bit of human effort was a lot simpler than trying to automate everything within the iframe to keep Turkers within the AMT environment. We implemented this and it worked perfectly.

On August 9, 23 weeks after we started development, we launched the discussion forum experiment and successfully collected data from 363 participants [7]. Now we were getting emails from Turkers that thanked us for the HIT that let them discuss important issues of the Turk experience, discussions that we later analyzed and summarized [5]. We also received an email from the TurkerNation community manager, inviting us to discuss and share the findings at their worker forum.

## LESSONS LEARNED

Our experience is a reminder that running studies with crowd workers is not a mere substitute for other ways of running user studies or behavioral experiments. Running a study in AMT meant we were not only researchers interacting with participants; we were also employers interacting with employees. Further, we learned how to face large volumes of simultaneous participants and receive support from participants and other researchers. But there were other lessons, too.

First, while the platform mediates the relationships between the researcher and the participant, it is designed for task-based work and not for research studies. This means that Requesters are allowed to reject uncompleted or low-quality work. Unfortunately, AMT is currently open for Turker abuse—Requesters can reject a Turker's work, not pay for it, and still use the data produced by the Turker. Besides these problems that make Turkers' work risky [5], researchers should remember that just as in traditional lab or field studies, participants should be compensated even if they do not complete the study properly or if they do not produce high-quality data.

Our experience is also a reminder that the people in the crowd are not simply *remote-human processors*. Beyond being an impressively powerful resource for user research [1], we found the Turkers we interacted with to be important collaborators, helping us debug the system, providing suggestions and advice on how to fix errors, and even offering social support. As such, our relationships with Turkers through the development of this study constantly shifted between employer-employees, researcher-participants, and researcher-collaborators, and we had to identify and be prepared for those shifts.

Second, the scale of participation is sometimes difficult to predict. Unlike traditional studies in the lab or in the field where the researcher controls the interaction with the participants, when working through a crowd work platform it is possible that all participants will show up almost at the same time to complete the study. This is good news for researchers, who can collect data from a user study with hundreds of participants in a short time. This also means that if the research system is outside the crowd-work platform, it has to be ready to support the load of participant activity.

More important, we learned that the experience of being a crowd researcher can be lonely at times, with one researcher facing sometimes dozens of crowd participants. Large volumes of aggressive emails are difficult to read, particularly while trying to track down system errors, and made us believe that there could be other frustrated Turkers who did not contact us. Fortunately, we found that although we were operating alone as a researcher-Requester, we received support and advice from Turkers and other experienced researchers. To make these resources more systematic, we call for better channels of communication in crowd platforms to more efficiently manage the communication—and potential collaboration—between crowd researchers and participants. Further, such tools could promote a community of practice for crowd-based user research that includes researchers in industry and academia, participants, and platform designers.

## ACKNOWLEDGMENTS

We thank all of the Turker participants for their contribution to our research and their support through system failures and recovery. Thanks to Six Silberman, Lilly Irani, Malte Jung, Jeff Hancock, and the Cornell IRB. This project is funded by NSF-HCC 1314778.

### ENDNOTES

1. Kittur, A., Chi, E., and Suh, B. Crowdsourcing user studies with Mechanical Turk. *Proc. of CHI '08*. 2008, 453–456.
2. McGinn, J. and LaRoche, C. Fast, cheap, and powerful user research. *Interactions 21*, 3 (May 2014), 62–65.
3. Mason, W. and Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods 44*, 1 (2012), 1–23.
4. Berinsky, A.J., Huber, G.A., and Lenz, G.S. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis 20* (2012), 351–368.
5. McInnis, B., Nam, C., Cosley, D., and Leshed, G. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. *Proc. of CHI '16*. 2016.
6. Irani, L.C. and Silberman, M.S. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. *Proc. of CHI '13*. 2013, 611–620.
7. McInnis, B., Murnane, E., Epstein, D., Cosley, D., and Leshed, G. One and done: Factors affecting one-time contributors to ad-hoc online communities. *Proc. of CSCW '16*. 2016, 609–623.

**Brian McInnis** is an information science Ph.D. student at Cornell University. His research focuses on the intersection between group dynamics and public policy in online spaces.
→ bjm277@cornell.edu

**Gilly Leshed** is a senior lecturer in information science at Cornell University. Her research focuses on how individuals and groups accomplish tasks and socialize, and the roles that technology plays in these interactions.
→ gl87@cornell.edu

DOI: 10.1145/2968077 