Advantages and Disadvantages of Remote Asynchronous Usability Testing Using Amazon Mechanical Turk

Erik T. Nelson & Angelos Stavrou
George Mason University

Amazon's Mechanical Turk is an online service that connects people all over the world who are willing to work for monetary compensation with companies and individuals who need simple tasks completed that are best done by humans. This paper examines the benefits and caveats from the use of Mechanical Turk as a platform to conduct an online usability test. Our findings indicate that Mechanical Turk can be a very useful tool when it is employed to collect basic user generated data such as click time and location, text input, and subjective usability data. In situations where more complex usage data is required, Mechanical Turk can still be a valuable resource to run a pilot test to assist in determining potential usability problems. However, in its current form, Mechanical Turk has limitations and can only be used to assess websites; it cannot be relied upon to conduct a usability test on software without the use of a virtual machine.

INTRODUCTION

Usability studies have been used for many years to test products, systems, and computer software. Historically, usability tests require that participants physically come to a lab where they are monitored while performing certain tasks. Recently, product developers have begun to test their software and websites using online test beds. Such online usability testing could have significant advantages. Compared to a traditional usability test, an online usability test has the potential to result in a significant reduction in the budget needed for the payment of participants. Participants who do not have to travel to a test site, but can instead participate in their own home, at their convenience, can be compensated less. In addition to the monetary savings for participant compensation, there is also a significant time savings for the experimenters. Rather than devoting multiple days to a single usability test with fewer participants, an experimenter can invest a few hours setting up the online test that will give him access to hundreds of participants in a relatively short amount of time. Online usability tests also have the potential to better reach participants from hard to include populations such as senior citizens (Tates et al., 2009). Even though senior citizens comprise a quickly growing group of Internet users, they are not as often included in usability tests because recruiting can be difficult. With an online usability test, senior citizens could be sent an invitation to participate from home.

While the online usability test has many advantages over a traditional usability test, there are several potential drawbacks that must be considered before proceeding with an online usability study. The lack of a controlled environment can make it difficult to determine if the subjects are distracted while completing the task. Moreover, self-selection, untruthful participants, and narrow populations can be other potential caveats of this methodology. Trust in web-based studies and questionnaires have long been an area of concern for researchers. While the validity of electronic questionnaires varies based on the nature of the topic in question, the general consensus is that it correlates very highly with the validity of questionnaires in a paper format (Smith & Leigh, 1997; Riva, Teruzzi, & Anolli, 2003) When it comes to usability testing, online tests also tend to correlate highly with their in-lab counterparts (Tullis, Fleischman, McNulty, Cianchette, & Bergel, 2002). Thompson, Rozanski, and Haake (2004) found that the same number of usability errors were detected using online and in-lab methods. While results on the efficacy of online usability tests vary, the general consensus is that this method provides valuable usability information, and should be considered.

It is also important that the purpose of the online usability test be understood. A usability test should be used as a means to identify and subsequently fix errors, inconsistencies, and potential problems that may confuse users. Conversely, some developers attempt to use usability tests to "prove" their software. While usability tests can never truly prove that an application is usable, with a large and diverse enough sample, it is possible that the application can be so thoroughly tested that it can generally be considered a usable product. An online usability test could definitely be used to identify errors, inconsistencies, and points of confusion regardless of whether the participants polled were distracted, untruthful, self-selected, or homogeneous. As long as potential errors are identified, the usability test has been successful. However, an online usability test is not appropriate for "proving" a website due to the potentially unrepresentative sample of users. This suggests that an online usability test may be more useful in the early stages of design.

PRACTICE INNOVATION

The idea of conducting a remote usability test is not a novel idea on its own. It began to receive attention in the mid-90s when Hammontree, Weiler, and Nayak (1994) published an article about the concept. As the concept has developed, two types of remote testing have emerged: synchronous and asynchronous.

Remote Synchronous Testing

With remote synchronous usability testing, the participant is in a different location than the experimenter, but the experimenter is still required to remotely administer the test. Oftentimes the experimenter is able to view the participant's screen using video capture software (Andreasen, Nielsen, Schroder, & Stage, 2007). This allows for very high fidelity data and can result in similar results to in-lab testing (Thompson et al., 2004). While these are encouraging findings, this method has two major drawbacks. The first drawback is that the experimenter's time is required when running every participant. This means that it will take just as long to complete a usability study as it would in a traditional lab-based study. The second drawback is that the location where the participant completes the experiment needs to be outfitted with recording equipment, including a video camera, a microphone, and screen recording software. This means that the participant cannot be depended upon to prepare the location and instead a representative from the usability team will likely need to travel to them.

Remote Asynchronous Testing

Asynchronous usability testing is the other type of remote usability testing. This differs from remote synchronous usability testing in that the experimenter is not required to be physically or virtually present while the participant is completing the study. While remote asynchronous testing has been conducted by researchers in several different ways, it typically falls into two major classes: auto logging and self-report by the participants (Brunn, Gull, Hofmeister, & Stage, 2009). With auto logging, information such as URLs visited, time spent on each page, links and buttons pressed, and other measures are automatically logged by the webpage. The experimenter then downloads and examines the data for usability errors. With self-reporting, the user completes each task, and then describes either how they interacted with the interface or what errors they ran into. The advantage of these methods is that the experimenters do not have to individually run each participant and participants can complete the study anywhere equipped with an Internet connection. This enables experimenters to collect data from many more participants than they normally would. On the downside, data is more limited and only 40% - 60% of errors identified using lab-based methods are normally detected given the same number of participants (Bruun et al., 2009). However, if an experimenter collected data from many more participants than they normally would by using remote asynchronous usability testing, it could be possible to produce similar results to an in-lab usability test. With Amazon's Mechanical Turk, collecting data from tens or even hundreds of participants is made possible.

Amazon's Mechanical Turk

Amazon's Mechanical Turk (www.mturk.com) is a website that was created as a marketplace for finding workers to do tasks that require human intelligence. It allows enterprises and regular users to gain access to a temporary online workforce rather than hiring a large temporary workforce to accomplish simple tasks that are best done by humans. Common tasks on Mechanical Turk include transcribing audio, identifying items in a photograph, or making sure a website doesn't have broken links to name a few. Companies and individuals, referred to as requesters, are able to create jobs, called Human Intelligence Tasks or HITs, for workers to accomplish, usually for minimal monetary compensation. One logical application of Mechanical Turk is to design HITs to remotely and asynchronously test the usability of a website. Workers, essentially usability testers, can then be compensated for successful completion of tasks. On Mechanical Turk, the range of compensation for the successful completion of a HIT is generally between $0.01 for a 5-10 minute HIT up to about $10.00 for a multi-hour or multi-day HIT, with the average HIT taking 10-20 minutes and paying between $0.02 and $0.10.

In our study, Mechanical Turk was evaluated as a possible tool for practitioners to use to conduct an online usability study. In addition, the effect of varying levels of compensation was studied.

PROCEDURE

For this evaluation, a HIT was designed on Mechanical Turk to test the usability of a website password interface. When a worker decided to work on this HIT, he or she was given a link to a website. Upon arrival at the website, participants completed a consent form and were then brought to the main experiment screen. Participants were next asked to complete a series of tasks, and were instructed to complete the tasks away from disruption, including music and televisions. After completion of the tasks, participants were given the System Usability Scale (SUS) questionnaire (Brooke, 1996) and were also asked for demographic information, including their technology

experience. The task took approximately 20-30 minutes, and participants were compensated $0.25, $0.50, $0.75, $1.00, $2.50, or $5.00. Blocks of HITs were created and remained open until 15 participants had participated at each level of compensation. After participants finished the study, their answers were submitted for review by the experimenters. All workers who completed the HIT were compensated. Any workers who did not complete the task had their HIT rejected, and did not receive compensation.

For this study, several types of data were collected. First, objective usability data, consisting of mouse click time and position, text entry, the current webpage being navigated, and navigation order, were recorded and saved to the website server. Finally, subjective data were collected using the SUS usability questionnaire as well as a demographics survey.

Note that in order to setup a website to work with this method of data collection, the website should first be hosted at a non-publically accessible web address. The user is then given a username and/or password to access the website. This ensures that the participant's interaction with the site is secure and can be traced back to their demographic and usability questionnaire data stored by Mechanical Turk. Lastly, the website must then be implemented in a way that enables the accurate recording of all user interactions (clicks, timing, current page, etc.) to a database. These are relatively easy modifications for an experienced information technology professional or computer scientist, and in our opinion, are well worth the upfront costs.

FINDINGS

One potential problem with using Mechanical Turk is that not every participant is focused in performing the task thoroughly. The participants' interest is geared towards quickly finishing the task. This can lead to some side effects: some users do not follow instructions correctly while others do not even attempt to follow the instructions, clicking through as quickly as they can. Although participants who do not complete the task do not need to be compensated, it still takes the experimenter time to sift through the measurements and sort out the valid from the invalid responses.

To explore the effect of compensation level, the total number of attempts per hour, number of valid attempts per hour and the percentage of valid to invalid attempts were calculated at all compensation levels. Not surprisingly, findings indicate that users of Mechanical Turk are highly motivated by increases in payment. Figure 1 shows the number of attempts per hour across all levels of payment. As would be expected, the more an experimenter is willing to pay, the more willing the participants are to participate. More interestingly, at a compensation level of $0.25, there was an average of only 0.075 attempts per hour the HIT was posted. In contrast, at a compensation level of $1.00, there was an average of 0.54 attempts per hour. This equals more than a 7 times increase in participants per hour for a 4 times increase in payment.

However, the number of participants who attempt to complete the task is much less important than the number of participants who follow the instructions for the task. Therefore, the ratio of valid attempts to invalid attempts was calculated for each compensation level, to see if the quality of participants' work changed with increased compensation. Figure 2 shows the proportion of valid to invalid responses at each level. At a compensation level of $0.25, participants had an average of 0.33 successes per attempt. The proportion of valid to invalid responses continued to increase until the compensation level of $2.50, after which it showed a marked drop. This indicates that, up to a certain point, participants try harder the more they are getting paid. This likely happens because they want to make sure that they do not have their HIT rejected so that they will receive payment. At the $5.00 compensation level, however, quality of performance dropped. This indicates the HIT process is governed by the law of diminishing returns. A compensation level of $2.50 is one of the higher paid HITs on the Mechanical Turk website. A compensation level of $5.00 gains a lot more attention from workers because it is significantly more than almost all other HITs available. At this compensation point, many people sign up to work on this HIT within the first few hours. However, many are just trying to complete the task as quickly as possible without following the instructions and hoping they will be paid anyway. Therefore, if one wanted to achieve the highest proportion of valid to invalid responses as quickly as possible, it is best to price a HIT higher than most other HITs. However, this value cannot be too much higher than the rest of the offered HITs to avoid attracting unnecessary attention.

While the purpose of this article is not to discuss the particular interface being evaluated, an interesting difference between the in-lab and Mechanical Turk methods surfaced. In this study, the password interface being studied required participants to click on very specific areas of an image, in the correct order, to correctly authenticate the password. We found that participants completing the task in the lab tended to painstakingly check and double-check the location before making a click. On the other hand, participants completing the task through Mechanical Turk tended to make clicks faster and less carefully, resulting in more failed authentications and a larger deviation between where the participant actually clicked compared to where they should have clicked. In the case of a password interface, normal users are very unlikely to spend a lot of time making sure they are pin-point accurate every time they enter their password. Perhaps the

lab environment caused in-lab participants to value accuracy over speed. This would make the results obtained from Mechanical Turk possibly more applicable to an actual user than the results obtained in the lab.

DISCUSSION

The current evaluation serves as a proof of concept that it is inexpensive, fast, and relatively easy to run a very basic online usability study for a website using Amazon's Mechanical Turk. In this evaluation, Mechanical Turk was used as a means of collecting objective and subjective usability data as well as recruiting and paying users.

The data in this study were collected by the experimental website. Therefore, only mouse click time and position, text entry, current webpage being navigated, and navigation order were collected. While this was sufficient for the current usability study, it may not be acceptable for more complex studies or websites. An alternative that would give the experimenter much more information, including a video of the participant's screen, would be to have participants log into a virtual machine running screen capture software hosted on a server, instead of just providing them a link to the website as demonstrated by Huang, Bias, Payne, and Rogers (2009). This method would have higher startup costs, but once the system was configured, it would be far superior to collecting data via the website alone. In fact, using a virtual machine, Mechanical Turk workers could participate in usability tests for computer software in addition to websites.

RECOMMENDATIONS

Testing Method

There are two approaches to collecting usability data using Mechanical Turk. The first is auto logging. This method is the only way to collect objective usability data using Mechanical Turk. However, the type of data that can be collected is limited to mouse click time and position, text entry, current webpage being navigated, and navigation order. These measures are appropriate for understanding how a person navigates through a website, but don't necessarily provide insight why they may be taking a less than ideal route. The other approach is participant self-report. While not objective, this method can be used to evaluate almost any website. Using this method, participants navigate a website and describe (in text) any pinch points that they encounter.

Participant Screening

Screening capabilities are relatively limited using Mechanical Turk. Currently, the screening options are limited to current location (country), HIT approval rate (how often the correctly complete HITs), and whether they have completed your HIT previously. While this doesn't allow for much sorting, you can make up for this by the sheer number of participants that you can compensate with a total investment of a few hundred dollars.

Costs & Time

For a task of medium complexity that takes about 20-30 minutes to complete, a compensation of $2.50 per participant seems to be the best at maximizing the quality of responses, while minimizing the time it takes to collect data. At this level of compensation, an experimenter should be able to collect approximately 5 valid responses overnight, 20 valid responses in about 3 days, and 100 valid responses in about 2 weeks.

Sensitivity

While Mechanical Turk is an excellent tool that allows for collecting large number of participants for little compensation, it is not recommended for new or prototype products or websites requiring non-disclosure agreements as they would be difficult if not impossible to enforce.

CONCLUSIONS

While an online usability study is not going to completely replace an in-lab study any time soon, it has the potential to be a very beneficial tool for usability professionals. Moreover, it can help weed out problems when a design is in its infancy and it allows for the quick testing of alternative prototype designs. In its current form, Mechanical Turk can serve as a very useful iterative diagnostic tool that can be employed to quickly collect usability data to help guide designers throughout the entire website or software design and implementation process.

ACKNOWLEDGMENTS

The authors would like to thank Kelley Baker and Sara Gee for their insightful reviews of this manuscript.

REFERENCES

Andreasen, M. S., Nielsen, H. V., Schroder, S. O., & Stage, J. (2007). What happened to remote usability testing? An empirical study of three methods. *CHI 2007 Proceedings*, 1405-1414.

Brooke, J. (1996) SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (eds.) Usability Evaluation in industry. London: Taylor and Francis.

Bruun, A., Gull, P., Hofmeister, L., & Stage, J. (2009). Let your users do the testing: A comparison of three remote asynchronous usability testing methods. *CHI 2009 Proceedings*, 1619-1628.

Huang, S. C., Bias, R. G., Payne, T. L., & Rogers, J. B. (2009). Remote usability testing: A practice. *JCDL Proceedings 2009.* 397.

Hammontree, M, Weiler, P., Nayak, N. (1994). Remote usability testing. *Interactions*, 1, 3, 21-25.

Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the internet in psychological research: comparison of online and offline questionnaires. *CyberPsychology and Behavior,* 6(1), 73-80.

Smith, M.A., & Leigh, B. (1997). Virtual subjects: using the internet as an alternative source of subjects and research environment. *Behavior Reserarch Methods, Instruments, & Computers,* 29(4), 496-505.

Tates, k., Zwaanswijk, M., Otten, R., van Dulmen, S., Hoogerbrugge, P.M., Kamps, W.A., & Bensing, J.M. (2009). Online focus groups as a tool to collect data in hard-to-include populations: examples from pediatric oncology. *BMC Medical Research Methodology,* 9:15, 1-8.

Thompson, K. E., Rozanski, E. P., & Haake, A. R. (2004). Here, there anywhere: Remote usability testing that works. *SIGITE Proceedings*, 132-137.

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., & Bergel, M. (2002). An empirical comparison of lab and remote usability testing of web sites. *Usability Professionals Association Conference Proceedings*.

FIGURES

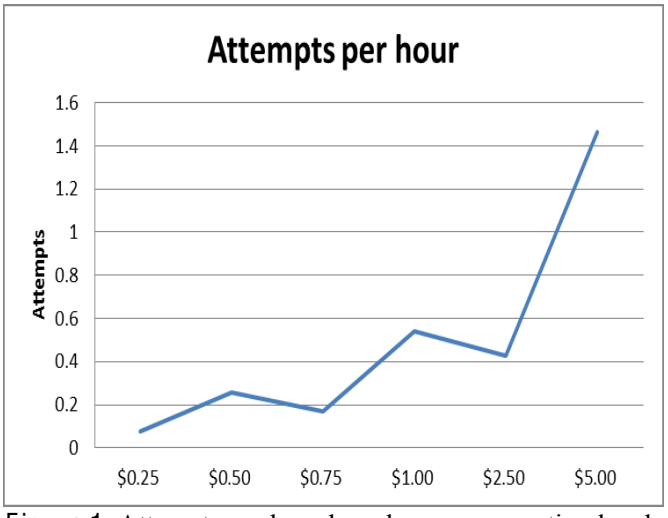

Figure 1. Attempts per hour based on compensation level.

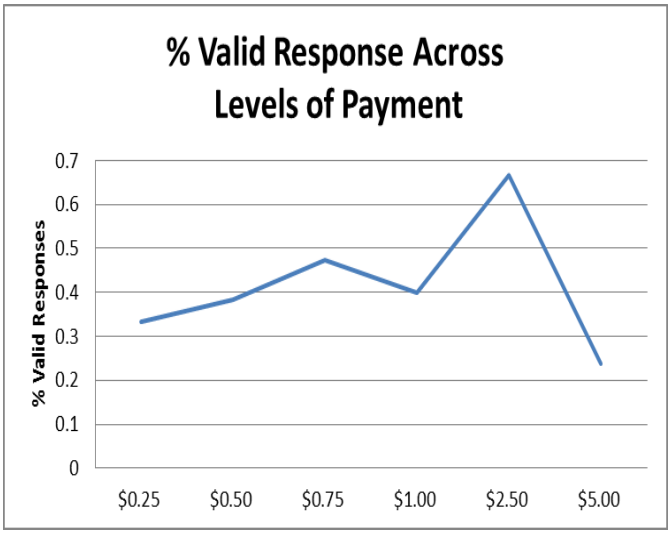

Figure 2. Ratio of valid to total number of responses based on compensation level.