# Crowdsourcing User Studies With Mechanical Turk

**Aniket Kittur, Ed H. Chi, Bongwon Suh**
Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304 USA
{nkittur, echi, suh}@parc.com

**ABSTRACT**
User studies are important for many aspects of the design process and involve techniques ranging from informal surveys to rigorous laboratory studies. However, the costs involved in engaging users often requires practitioners to trade off between sample size, time requirements, and monetary costs. Micro-task markets, such as Amazon's Mechanical Turk, offer a potential paradigm for engaging a large number of users for low time and monetary costs. Here we investigate the utility of a micro-task market for collecting user measurements, and discuss design considerations for developing remote micro user evaluation tasks. Although micro-task markets have great potential for rapidly collecting user measurements at low costs, we found that special care is needed in formulating tasks in order to harness the capabilities of the approach.

**Author Keywords**
Remote user study, Mechanical Turk, micro task, Wikipedia.

**ACM Classification Keywords**
H5.2 Information interfaces and presentation (e.g., HCI) --- User Interfaces: Evaluation/methodology, Theory and Methods; H.5.3 [Information Interfaces]: Group and Organization Interfaces –Web-based interaction.

**INTRODUCTION**
User studies are vital to the success of virtually any design endeavor. Early user input can substantially improve the interaction design, and input after development can provide important feedback for continued improvement. User evaluations may include methods such as surveys, usability tests, rapid prototyping, cognitive walkthroughs, quantitative ratings, and performance measures.

An important factor in planning user evaluation is the economics of collecting user input. There are monetary costs of acquiring participants, observers, and equipment; in addition, some techniques are more time intensive than others. Thus it is often not possible to acquire user input that is both low-cost and timely enough to impact development. The high costs of sampling additional users lead practitioners to trade off the number of participants with monetary and time costs [5].

Collecting input from only a small set of participants is problematic in many design situations. In usability testing, many issues and errors (even large ones) are not easily caught with a small number of participants [5]. In both prototyping and system validation, small samples often lead to a lack of statistical reliability, making it difficult to determine whether one approach is more effective than another. The lack of statistical rigor associated with small sample sizes is also problematic for both experimental and observational research.

These factors have led to new ways for practitioners to collect input from users on the Web, including tools for user surveys (e.g., surveymonkey.com, vividence.com), online experiments [3], and remote usability testing [2]. Such tools expand the potential user pool to anyone connected to the internet. However, many of these approaches still either rely on the practitioner to actually recruit participants, or have a limited pool of users to draw on.

In this article we investigate a different paradigm for collecting user input: the micro-task market. We define a micro-task market as a system in which small tasks (typically on the order of minutes or even seconds) are entered into a common system in which users can select and complete them for some reward which can be monetary or non-monetary (e.g., reputation). Micro-task markets offer the practitioner a way to quickly access a large user pool, collect data, and compensate users with micro-payments. Here we examine the utility of a general purpose micro-task market, Amazon's Mechanical Turk (mturk.com), as a way to rapidly collect user input at low cost.

### Micro-task Markets: Mechanical Turk
Amazon's Mechanical Turk is a market in which anyone can post tasks to be completed and specify prices paid for completing them. The inspiration of the system was to have human users complete simple tasks that would otherwise be extremely difficult (if not impossible) for computers to perform. A number of businesses use Mechanical Turk to source thousands of micro-tasks that require human intelligence, for example to identify objects in images, find relevant information, or to do natural language processing.


CHI 2008, April 5–10, 2008, Florence, Italy.

Tasks typically require little time and effort, and users are paid a very small amount upon completion (often on the order of a few cents). In March 2007, Amazon claimed the user base of Mechanical Turk (who commonly refer to themselves as "turkers") consisted of over 100,000 users from over 100 countries[1].

Adapting this system for use as a research and design platform presents serious challenges. First, an important driver of the success of the system appears to be the low participation costs for accepting and completing simple, short tasks [1]. In contrast, paradigms for user evaluation traditionally employ far fewer users with more complex tasks, which incur higher participation costs.

Second, Mechanical Turk is best suited for tasks in which there is a bona fide answer, as otherwise users would be able to "game" the system and provide nonsense answers in order to decrease their time spent and thus increase their rate of pay. However, when collecting user ratings and opinions there is often no single definite answer, making it difficult to identify answers provided by malicious users.

Third, the diversity and unknown nature of the Mechanical Turk user base is both a benefit and a drawback. Since many users are sampled with a pool drawn from all over the globe, results found using the Mechanical Turk population have the potential to generalize to a varied population more than the small user samples and limited geographic diversity typical of more traditional recruiting methods. On the other hand, the lack of demographic information, unknown expertise, and limited experimenter contact with the Mechanical Turk population raise the question of whether useful data can be collected using micro-task markets.

## EXPERIMENTS

We conducted two experiments to test the utility of Mechanical Turk as a user study platform. We used tasks that collected quantitative user ratings as well as qualitative feedback regarding the quality of Wikipedia articles.

Wikipedia is an online encyclopedia which allows any user to contribute and change content, with those changes immediately visible to visiting users. Assessing the quality of an article in Wikipedia has been the subject of much effort both from researchers [4] and the Wikipedia community itself [7]. A rapid, robust, and cost-effective method for assessing the quality of content could be useful for many other systems as well in which content is contributed or changed by users.

We conducted an experiment in which we had Mechanical Turk users rate a set of 14 Wikipedia articles, and then compared their ratings to an expert group of Wikipedia administrators from a previous experiment [4]. Admins are highly experienced Wikipedia users with a strong track record of participation. Articles were originally chosen for a different purpose (randomly sampled to examine a range of degrees of conflict), but included a range of expert-rated quality ratings. Old versions of the articles (from 7/2/2006) were used so that turkers would see the same content as the admins. Examples of articles include "Germany", "Noam Chomsky", "Hinduism", and "KaDee Strickland", amongst others.

### Experiment 1

In the first experiment we attempted to mirror the task given to admins as closely as possible. Thus, similar to the original admin task, we had users rate articles on a 7-point Likert-scale according to a set of factors including how well written, factually accurate, neutral, well structured, and overall high quality the article was. These questions were taken from the Wikipedia "Featured article criteria" page as guidelines for vetting high-quality articles [7]. Brief descriptions of what was meant by each question were presented as part of the question, again summarized from the Wikipedia featured article criteria.

In addition, users were required to fill out a free-form text box describing what improvements they thought the article needed. This was done primarily to provide a check on whether users had in fact attended to the article or had just provided random ratings. Turkers were paid 5 cents for each task completed.

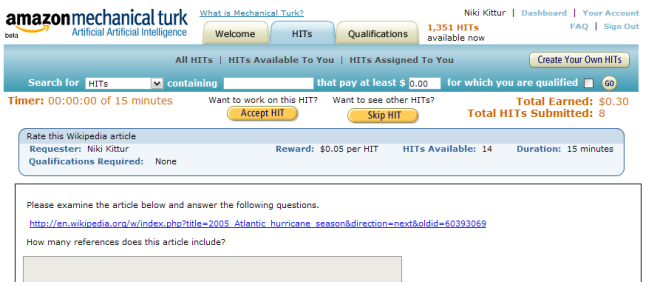


**Figure 1. Amazon's Mechanical Turk (mturk.com) allows users to preview a task, see the payment offered, and how many instances remain available.**

#### *Results*

58 users provided 210 ratings for 14 articles (i.e., 15 ratings per article). User response was extremely fast, with 93 of the ratings received in the first 24 hours after the task was posted, and the remaining 117 received in the next 24 hours. Many tasks were completed within minutes of entry into the system, attesting to the rapid speed of user testing capable with Mechanical Turk.

However, the correlation between Mechanical Turk user ratings and Wikipedia admin ratings was only marginally significant ($r$ = 0.50, $p$ = .07), providing only very weak support for the utility of Mechanical Turk ratings mirroring expert ratings. Furthermore, a closer look at user responses suggested widespread "gaming" of the system. Out of the total of 210 free-text responses regarding how the article could be improved, 102 (48.6%) consisted of uninformative

[1] http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html

responses including semantically empty (e.g., "None"), non-constructive (e.g., "well written"), or copy-and-paste responses (e.g., "More pictures to break up the text" given for all articles rated by a user). An examination of the time taken to complete each rating also suggested gaming, with 64 ratings completed in less than 1 minute (less time than likely needed for reading the article, let along rating it). 123 (58.6%) ratings were flagged as potentially invalid based either on their comments or duration. The remaining responses were too sparse to conduct a robust statistical analysis.

However, many of the invalid responses were due to a small minority of users. Only 8 users gave 5 or more responses flagged as potentially invalid based on either comments or time; yet these same users accounted for 73% (90 responses) of all flagged responses. Thus it appeared that, rather than widespread gaming, a small group of users were trying to take advantage of the system multiple times.

### Experiment 2

The results from Experiment 1 provided only weak support for the utility of Mechanical Turk as a user measurement tool. Furthermore, they demonstrated the susceptibility of the system to malicious user behavior. Even though these users were not rewarded (invalid responses were rejected), they consumed experimenter resources in finding, removing, and rejecting their responses.

In experiment 2, we tried a different method of collecting user responses in order to see whether the match to expert user responses could be improved and the number of invalid responses reduced. The new design was intended to make creating believable invalid responses as effortful as completing the task in good faith. The task was also designed such that completing the known and verifiable portions of the questionnaire would likely give the user sufficient familiarity with the content to accurately complete the subjective portion (the quality rating).

All procedures were identical to Experiment 1, except that the rating task was altered. In the new rating task, users were required to complete four questions that had verifiable, quantitative answers before rating the quality of the article. Questions were selected to remain quantitative and verifiable yet require users to attend to similar criteria as the Wikipedia featured article criteria, and as what Wikipedia administrators claimed they used when rating an article. These questions required users to input how many references, images, and sections the article had. In addition, users were required to provide 4-6 keywords that would give someone a good summary of the contents of the article. This question was added to require users to process the content of the article as well as simply counting various features, while being more objective and verifiable than the request for constructive feedback in Experiment 1. Users were then asked to provide a rating of the overall quality of the article, described as "By quality we mean that it is well written, factually comprehensive and accurate, fair and without bias, well structured and organized, etc.", again on a 7-point Likert scale. Finally, users were also asked in a free-text field to give as much insight as possible into their decision.

| | Invalid Comment Responses | Median duration | Duration < 1 minute |
|---|---|---|---|
| Exp 1 | 48.6% | 1:30 | 30.5% |
| Exp 2 | 2.5% | 4:06 | 6.5% |

**Table 1. Improvement in response quality in Experiment 2 upon the introduction of verifiable questions.**

#### *Results*

124 users provided 277 ratings for 14 articles (i.e., 19-20 ratings per article). The number of ratings per user was significantly smaller than for Experiment 1 (2.2 vs. 3.6; $t(180) = 2.84$, $p < .01$) and also more distributed across users (only 5% of users rated 8 or more pages, vs. 16% in Experiment 1). The positive correlation between Mechanical Turk and Wikipedia administrator ratings was also higher than Experiment 1, and was statistically significant ($r = 0.66$, $p = 0.01$).

In addition to the improved match to expert ratings, there were dramatically fewer responses that appeared invalid. Only 7 responses had meaningless, incorrect, or copy-and-paste summaries, versus 102 in Experiment 1. Also, only 18 responses were completed in less than one minute, and the median completion time was much higher than in Experiment 1 (4:06 vs. 1:30; $t(486) = 5.14$, $p < .001$).

## DISCUSSION

In Experiment 1 we found only a marginal correlation of turkers' quality ratings with expert admins, and also encountered a high proportion of suspect ratings. However, a simple redesign of the task in Experiment 2 resulted in a better match to expert ratings, a dramatic decrease in suspect responses, and an increase in time-on-task.

The match to expert ratings is somewhat remarkable given the major differences between the turkers and the admins. Since the turker population is drawn from a wide range of users, they represent a more novice perspective and likely weight different criteria in making quality judgments than the highly expert admin population. The correlation between the two populations supports the utility of using crowds to approximate expert judgments in this setting. For some applications in which collecting many varied data points is important, such as prototype testing or user measurements, judgments from a varied crowd population may be even more useful than a limited pool of experts.

#### *Design Recommendations*

The strong difference between the two experiments points to design recommendations for practitioners looking to harness the capabilities of micro-task markets:

First, it is extremely important to have explicitly verifiable questions as part of the task. In Experiment 2 the first four questions users answered could be concretely verified. Not all of these questions need to be quantitative; one of the most

useful questions turned out to be asking users to generate keyword tags for the content, as the tags could be vetted for relevance and also required users to process the content. Another important role of verifiable questions is in signaling to users that their answers will be scrutinized, which may play a role in both reducing invalid responses and increasing time-on-task.

Second, it is advantageous to design the task such that completing it accurately and in good faith requires as much or less effort than non-obvious random or malicious completion. Part of the reason that user ratings in Experiment 2 matched up with expert ratings more closely is likely due to the task mirroring some of the evaluations that experts make, such as examining references and article structure. These tasks and the summarization activity of keyword tagging raise the cost of generating non-obvious malicious responses to at least as high as producing good-faith responses.

Third, it is useful to have multiple ways to detect suspect responses. Even for highly subjective responses there are certain patterns that in combination can indicate a response is suspect. For example, extremely short task durations and comments that are repeated verbatim across multiple tasks are indicators of suspect edits.

### *Advantages and Limitations*

In this study we examined a single user task using Mechanical Turk, finding that even for a subjective task the use of task-relevant, verifiable questions led to consistent answers that matched expert judgments. These results suggest that micro-task markets may be useful for other types of user study tasks that combine objective and subjective information gathering. For example, Mechanical Turk could be used for rapid iterative prototyping by asking users a number of verifiable questions regarding the content and design of a prototype followed by a subjective rating; or for surveying users by asking them to fill out common-knowledge questions before asking for their opinion; or for online experiments by collecting objective measurements prior to subjective responses.

However, Mechanical Turk also has a number of limitations. Some of these are common to online experimentation: for example, ecological validity cannot be guaranteed, since there is no easy way for experimenters to fully control the experimental setting, leading to potential issues such as different browser experiences or distractions in the physical environment. Moreover, Mechanical Turk does not have robust support for participant assignment, making even simple between-subject designs difficult to execute. However, there is support for qualifying users by using automated pre-tests, or for including or excluding users from future tasks based on their responses to past tasks.

It is possible to simply use Mechanical Turk as a recruitment device and to host the user study oneself using a simple API to send and receive participant information from Amazon. In this case the restrictions on participant assignment are removed as all the work is done on the experimenters' side; however, this also requires significantly more programming and setup resources to execute.

Further work is needed to understand the kinds of experiments that are well-suited to user testing via micro-task markets and determining effective techniques for promoting useful user participation. For example, one research question is whether participants might police each other [1] in micro-task markets. Also, tasks requiring significant interaction between users (for example, collaboratively creating content) might be less suitable for using a micro-task market than independent tasks. Given the many advantages of micro-task markets, understanding the types of tasks they are effective for is an important area for future research.

## CONCLUSION

Micro-task markets such as Amazon's Mechanical Turk are promising platforms for conducting a variety of user study tasks, ranging from surveys to rapid prototyping to quantitative performance measures. Hundreds of users can be recruited for highly interactive tasks for marginal costs within a timeframe of days or even minutes. However, special care must be taken in the design of the task, especially for user measurements that are subjective or qualitative.

## REFERENCES


1. Benkler, Y. 2002. Coase's penguin,or Linux and the nature of the firm. *Yale Law Journal* 112, 367-445.
2. Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., and Stage, J. 2007. What happened to remote usability testing? An empirical study of three methods. In *Proc. of CHI 2007*. ACM Press, New York, NY, 1405-1414.
3. Fogg, B.J., Marshall, J., Kameda, T., Solomon, J., Rangnekar, A., Boyd, J., and Brown, B. 2001. Web credibility research: a method for online experiments and early study results. In *Proc. CHI '01 Extended Abstracts.* ACM Press, New York, NY, 295-296.
4. Kittur, A., Suh, B., Pendleton, B., Chi, E.H.. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proc. of CHI 2007*, pp. 453--462. ACM Press.
5. Spool, J, & Schroeder, W. 2001. Test web sites: five users is nowhere near enough. In *Proc. CHI2001 Extended Abstracts*. Seattle: ACM Press. 285-286.
6. Viégas, F.B., Wattenberg, M., and McKeon, M. 2007. The Hidden Order of Wikipedia. In *Proc. of HCI Interactional Conference*.
7. Wikipedia. Featured Article Criteria. http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria, accessed Sep, 2007.