Investigating the Efficacy of Crowdsourcing on Evaluating Visual Decision Supporting System

Crowdsourcing recently became a popular approach to substitute time consuming and expensive human subject studies, but its application is generally limited to simple and short-term experimental tasks, such as testing visual perception. The goal of this study is to test if crowdsourcing is applicable to a more complicated user study. Thus, we replicated a controlled lab study of decision-making tasks with different sorting techniques using crowdsourcing. Total 98 participants were recruited via the Amazon Mechanical Turk service, and they participated in the study remotely through web interfaces. Results of the experiment indicate that crowdsourcing experiment is not exactly equivalent to lab experiments. However, we found potential sources of problems that we can improve to make the crowdsourcing experiment more viable.

INTRODUCTION

Controlled laboratory studies are one of core methods to test the efficacy of various user interfaces. Despite their disadvantages in sacrificing external validity, they are widely adopted due to its benefits of the full control over environments and participants, which increase internal validity. However, controlled laboratory studies with college students often lack representativeness of the whole user population, and it often costs too much money to recruit lots of participants. Web-based experiments could resolve some of these problems (Reips, 2002), but it is still challenging to recruit a large number of participants promptly and to build trust between experimenters and participants.

Crowdsourcing seems to be an ideal solution to these problems. It provides an instant access to a large and diverse pool of participants at less cost, and crowdsourcing platforms provide billing mechanisms that participants and experimenters can trust. Amazon Mechanical Turk (MTurk) is one of the leading crowdsourcing platforms, which is getting more popularity among researchers.

However, we realized some potential drawbacks in using MTurk as a platform for rather complicated user studies. Majority of tasks posted in MTurk are used for surveys, image labeling, and natural language processing, which are often relatively simple and short-term tasks (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). Therefore, when more complicated tasks are posted, MTurk workers (Turkers) might not be interested or may leave the task incomplete, which deteriorates data quality. Thus, before adopting crowdsourcing as an alternative way to conduct user studies with burdensome cognitive processes, it is necessary to test if we can collect quality data through the system.

Therefore, we replicated a complicated user study in MTurk to test the viability of crowdsourcing. The study involves multivariate object selection tasks using different visualization techniques. In the following sections, we describe the background for MTurk and the used visualization techniques. Then, we discuss the method of our experiment. At the end, we present our findings in results and discussions with future work.

BACKGROUND

Mechanical Turk

As a convenient and efficient platform to recruit a large number of workers for small tasks, MTurk has been adopted as a platform for human subject studies. Studies using MTurk focus on natural language processing micro tasks, such as transcription of spoken language data (e.g., Marge, Banerjee, & Rudnicky, 2010; Bloodgood & Callison-Burch, 2010), finding word sense disambiguation (e.g., Akkaya et al., 2010), and annotation tasks on headlines and images to explore the emotion it contains (e.g., Snow et al., 2008; Sorokin & Forsyth, 2008). Likewise, the inspiration of using MTurk in research studies was to complete simple tasks, which could be done easily with human users but difficult with computer machines.

Previous research on MTurk indicated researchers were always suspicious about the validity of the system due to some inherent disadvantages. In order to assess the validity of MTurk, Heer and Bostock (2010) replicated previous lab experiments and found that the crowdsourced results were consistent with prior findings. Paolacci et al. (2010) also compared three classic experimental tasks for decision making in certain scenarios drawn from the heuristics and biases literature. Though the results from MTurk showed that people tended to be more risk aversive than the traditional subject pool, the overall data showed the similar trend as the previous results. Based on the literature, we organized weaknesses of crowdsourcing as an experimental platform:

Lack of Control. In order to collect reliable data, participants are usually required to concentrate on the task in an uninterrupted environment. However, researchers do not have the control over how seriously Turkers participate in MTurk experiments (Marge, Banerjee, & Rudnicky, 2010; Bloodgood & Callison-Burch, 2010; and Stolee & Elbaum, 2010).

Difficult Screening Process. Though MTurk provided some screening procedures, such as filtering based on demographical information or experiences as a Turker, it is still challenging to have a well-balanced mixture of participants (Sala, Partridge, Jacobson & Begole, 2007; Kelley, 2010 and Stolee & Elbaum, 2010).

Uncertain Data Quality. We cannot guarantee the data quality due to the low payment and the anonymity of Internet. Turkers usually have little motivation to work on the task (Paolacci, Chandler, & Ipeirotis, 2010; Mason & Watts, 2009; Kosara, & Ziemkiewicz, 2010).

Demographic Issue. The recent study on the Mechanical Turkers indicated that Mechanical Turkers are not precisely representative of the U.S. population, and the homogeneity of Turkers' education levels may be a potential problem for some research design (Paolacci, Chandler, & Ipeirotis, 2010; Ross et al., 2009).

Visualized Decision Making

Despite these drawbacks, we found that crowdsourcing is very attractive to conduct user studies especially for testing information visualization systems for decision making, which we call "visualized decision making (VDM)." Since decision-making activities are subject to various factors (e.g., the number of choices/attributes, the presence and absence of different interaction/visualization techniques, and individual differences), conducting studies with off-line participants are too time-consuming and expensive.

One of VDM techniques, called SS, was proposed as an interactive table that sorts multiple attributes at the same time to support compensatory decision tasks (A, 2009)[1]. In SS, an item is not presented in a single row. Instead, SS visualizes a cell at a position where the vertical position means the rank of the item in the corresponding attribute. The higher a cell value is positioned, the higher the item's value in the corresponding attribute is. So, the positions of cells belonging to a single item allows a user to easily identify the overview values of an item in multiple attributes. In the following study (A, 2010), we also created another variation of SS called PT, which positions items in the faithful vertical position.

A (2010) tested four different sorting techniques: SS, PT, a table with the typical sorting feature (TS) and a static table without a soring feature (B). A task used in their study was to select the highest value items (sum of values across attributes). We incentivized the participants with higher payment as they select items more accurately and quickly. The results of controlled lab studies indicated that participants using SS and PT demonstrated higher decision qualities than those using TS and B even using less time.

METHODS

We replicated our previous controlled laboratory experiment (A, 2010) using MTurk. We compared four sorting techniques with the same task, a multi-attribute object selection task, with the identical data. Four sorting techniques were tested, TS, SS, PT, and B. Since we could not post the SS system on the standard MTurk website, we separately hosted our experimental web system, developed using Ruby on Rails (http://rubyonrails.org/). The four sorting techniques are implemented using Flex (http://www.adobe.com/products/flex) and Flare (http://flare.prefuse.org/) so that participants can access via web. Figure 1 shows the experimental interface with the SS treatment.

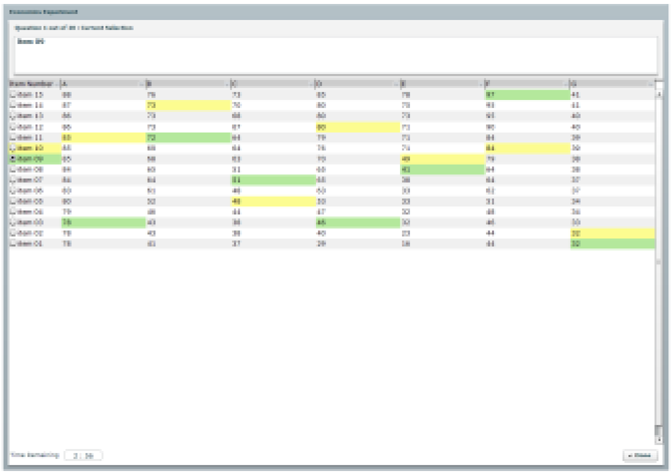

Figure 1. Screen shot of the interface with SS. All the columns are sorted simultaneously where the highlighted color corresponds to one item.

After the participants read the instructions on the standard MTurk website, they were redirected to our system with a unique login ID. After the experiment was done, they were given a finish code. We could also collect other miscellaneous data, such as display configuration from the participants, in order to maintain the most identical environment between Turkers. To compare the results we kept the experiment design as same as possible to the lab experiment.

Participants

As the previous lab experiment recruited 20 participants per treatment, we collected data from 30 participants assuming that there would be Turkers who quit during the experiment. Among the 120 participants, 12 (10%) dropped out during the experiment, and 10 (9%) were sifted out in the middle of a trial. Overall, we collected 98 (82% of the all the participants) valid responses.

Procedures

After the participants were redirected to our system, each participant was asked to complete 20 trials of tasks. Each participant was given one of the four interfaces using different sorting techniques (B, TS, SS, and PT). Instructions were given to verify the participants' understanding about the tool and the experiment procedure.

Task

The task was to select an item with the highest-value out of 15 items with 7 attributes. The value of an item is calculated by the sum of its normalized attribute values as in the following equation:

[1] The citation was removed for blind review

$$value_i = \sum_{j=0}^{7} \frac{T_{ij} - \min T_{.j}}{\max T_{.j} - \min T_{.j}} \qquad (1)$$

value$_i$ is the i^{th} item's value; T_{ij} is the number in the j^{th} column (attribute) of the i^{th} row (item) in data set *T*. Basically, *value*$_i$ is the summation of attribute values, each of which is normalized within each column. Though this equation looks arbitrary, it was designed to avoid various gaming (A, 2010) and actually resembles a real life decision-making when a person consider multiple attributes equally. Participants had 20 trials to select the highest-value item within 3-minute time limit. We used different dataset for each trial.

Incentives

After each trial, a participant earned payment by (A) the value of the finally selected item plus (B) the value of an item selected at a randomly selected time. Compensation (B) was designed to promote participants to select the best choice at any given time, which is inspired by (Caplin et al., n.d.). In addition, we also paid twenty-cent show-up compensation. On average, each participant earned $0.72 for participating, but compensation for each participant varies ($0.45 - $1.02), depending on his or her performance. This is lower than that those of the previous lab experiment (approximately $20 per participation). Another difference is that we compensated participants based on three randomly selected trials out of 20 trials in order to make each trial financially substantial ($ more than $6 per trial) in our previous lab experiment, but we compensated Turkers smaller wages for every single trial in order to conform to the commonly accepted wage structure at MTurk: a small wage for each small task. We believe that this approach is less confusing to Turkers and actually motivates them to participate consistently without losing concentration throughout the 20 trials. Each trial earning was approximately between 4 cents and 10 cents based on the value of the item selected.

RESULTS[2]

Demographic Summary

Among the 98 valid responses, 30 were female. The average age of the participants was 28.9, ranging from 18 to 54. The overall educational level of the participants was relatively higher (some have even professional degrees), where the lab study participants were mostly all undergraduate students.

Efficiency

We replicated A's data analysis for decision quality using the *efficiency* measure because the highest value of each item changes depending on the dataset. It is calculated as below:

$$efficiency_i = \frac{value_i - \min(value.)}{\max(value.) - \min(value.)} \qquad (2)$$

value$_i$ is calculated from Equation 1 where 0.00 is the lowest efficiency and 1.00 is the highest efficiency for an item.

The MTurk data results do not show statistically significant difference in efficiency between four sorting techniques (F(3,94) = .90, p = .44). Thus, the result was not consistent with that of the lab experiment where SS and PT both had significantly higher efficiency compared to B and TS.

We found significant difference in efficiency of each sorting technique between the lab experiment and the Mturk experiment. We compared mean and standard deviation between two studies in each sorting technique. The mean efficiency of all sorting techniques in the lab experiment was significantly higher than that in the Mturk experiment: B (z = 10.77, p < 0.0001), TS (z = 9.33, p < 0.0001), SS (z = 11.88, p < .0001), PT (z = 15.03, p < .0001). The average efficiency across all treatments was 0.89 for the lab experiment and 0.66 for the MTurk study. Figure 2 summarizes average efficiency and standard deviation for four sorting techniques in the lab experiment (blue) and the MTurk experiment (red). The general trend shows that the efficiency of SS was the highest and that of B is the lowest in both studies.

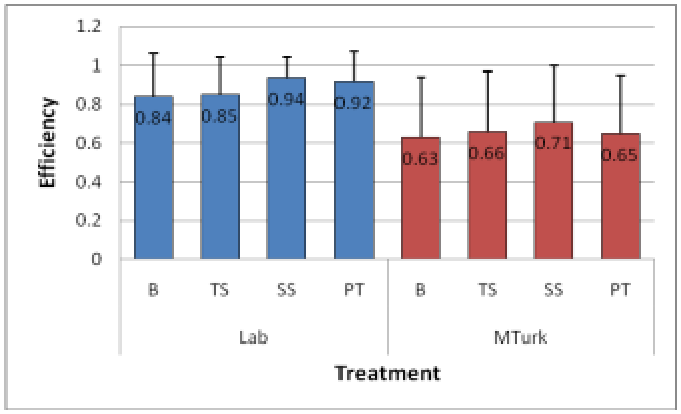

Figure 2. Bar graph of the mean and standard deviation of efficiency for each treatments.

Time Spent

There was no significant difference in time spent between participants in the MTurk experiment (F(3,94) = 1.84, p = 0.14). On the other hand, TS spent longer time in average than the other sorting techniques in the lab experiment. We found significant difference in time spent of each sorting technique between the lab experiment and the MTurk experiment. We compared mean and standard deviation between two studies in each sorting technique. The mean time spent of all sorting techniques in the lab experiment was significantly higher than that in the MTurk experiment: B (z = 19.95, p < .0001), TS (z = 18.38, p < .0001), SS (z = 18.14, p < .0001), PT (z = 21.41, p < .0001). MTurk participants spent 25.7 seconds in average for each trial. On the other hand, the lab participants spent 83.8 seconds in average. Figure 3 summarizes the average time spent and standard deviation for four sorting techniques in the lab experiment and the MTurk experiment. The general trend shows that participants in both studies spent more time on TS than other sorting techniques.

[2] Since we are comparing the result from this study and that from the previous study, we use "the MTurk experiment" for this study and "the lab experiment" for the previous study for convenience.

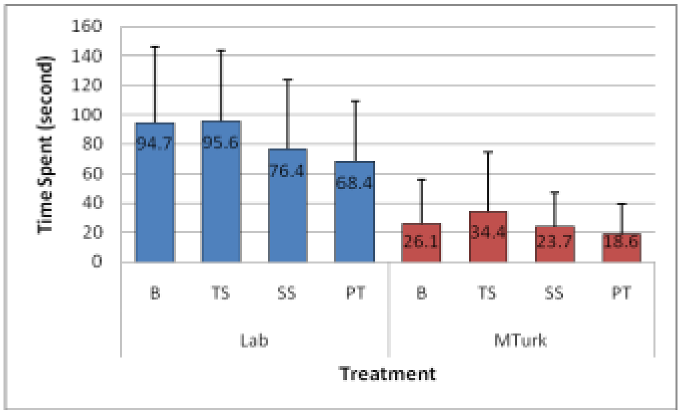

Figure 3. Bar graph of the mean and standard deviation of time spent for each treatment.

DISCUSSION

From our initial study, we cannot conclude that MTurk can be readily used to conduct a decision making experiment. We failed to find consistent results between the MTurk experiment and the lab experiment. However, we also observed that the general trends in time spend and efficiency were partially preserved in the MTurk experiment. This could indicate that we have room to improve.

We conjecture that the main problem of MTurk is less control over the participants. In Figure 3, average time spent in the MTurk study is significantly lower than that in the lab experiment. To analyze the phenomena more deeply, we plotted the distribution of time spent in Figure 4. We found that there is a high peak around 5 seconds from the beginning. We believe this indicates that some participants select items randomly to quickly finish the experiment. Similarly, we could also observe more fluctuating efficiency across 20 trials from MTurk participants in Figure 5. These “bad Turkers” may pollute the results from “faithful Turkers.” To minimize the impact of this issue, we should consider a better experiment design to attract the faithful Turkers or promote Turkers to respond seriously. At the same time, we should implement a rigorous screening process to detect the bad Turkers who game the system to earn money without effort.

Based on the lessons we learned, we propose how we would change the system in the future:

Adjust the payment system and time structure. We assume that the main problem with our payment system is that Turkers can earn reasonable amounts of money by randomly selecting items. With our payment structure, actually spending more time to choose a better option is not an economically rational behavior. It is much better to finish up tasks quickly if potential gain is not substantial. While preventing this behavior, we also need a systematic way to encourage them to do the task properly. For our experiment one way is to pay only for the rounds where they select the top three ranked items or apply an exponential curve to weight the items that rank higher. We also observed that some people were eager to click through the trials. To avoid this, we could prevent them from escaping from or moving away from the screen during the full 3-minute per each trial. We do not know how Turkers will respond to this enforcement, however, as it is a fairly long time with 20 rounds, we believe that the Turkers who remain are willing to spend the time doing the task properly.

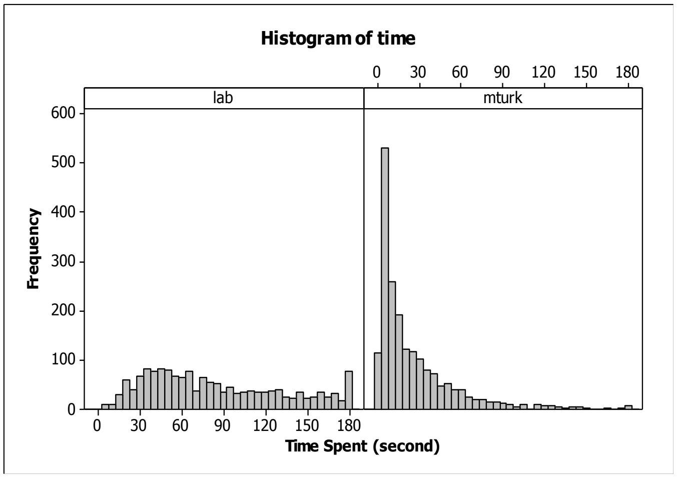

Figure 4. Histogram on time spent for the lab experiment and the MTurk experiment.

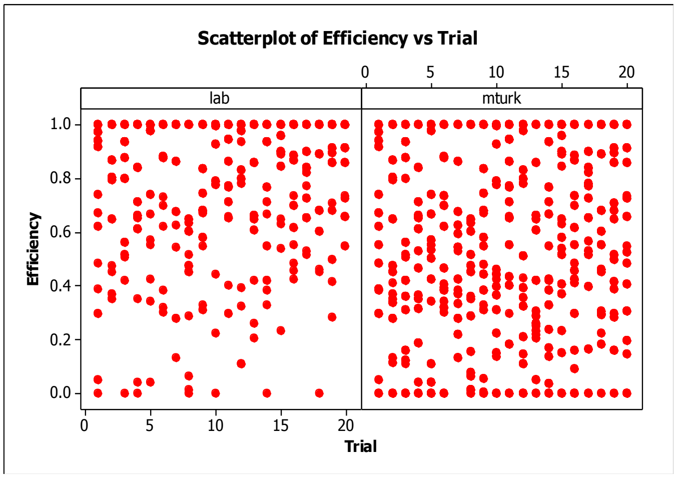

Figure 5. Scatterplot on efficiency for the lab experiment and the MTurk experiment.

Proper screening process. Even though we carefully designed the experiment to prevent “bad Turkers,” it is difficult to avoid this issue. After collecting the data, there needs to be a way to screen out the useless data. To make the screening process work, the criteria should be embedded in and related to the task itself (Snow, O’Connor, Jurafsky & Ng, 2008). We attempted to use the pre-quizzes as a qualification method. However, even though they solved it properly, it did not guarantee that they performed faithfully throughout all the trials. If we could define good indicators of faithfulness, these could help refine data.

CONCLUSION

We cannot draw a conclusion about the validity of MTurk as a platform for cognitively burdensome tasks, such as decision-making using visualization techniques. We merely identified potential challenges in conducting crowdsourcing studies and summarized our lessons for future use. However, we could recruit 98 participants within two days, which proves, "hundreds of users can be recruited for highly interactive tasks for marginal costs within a timeframe of days or even minutes" (Kittur, Chi, & Suh, 2008). To improve the MTurk experiment, we need to carefully design the small details of experiments (e.g., payment structure and instruction) to recruit more “faithful Turkers.”

REFERENCES

A (2009). Hidden for blind review.

A (2010). Hidden for blind review.

Akkaya, C., Conrad, A., Wiebe, J., & Mihalcea, R. (2010). Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010*. Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, Los Angeles. 195–203.

Bloodgood, M., & Callison-Burch, C. (2010). Using mechanical turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language* Data with Amazon’s Mechanical Turk, Los Angeles. 208–211.

Caplin, A., Dean, M., & Martin, D. (n.d.). Search and Satisficing! Working Paper.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*. 203–212.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. 453-456.

Kosara, R., & Ziemkiewicz, C. (2010). Do Mechanical Turks Dream of Square Pie Charts? *In BELIV ’10: Beyond Time and Errors: Novel Evaluation Methods for Information Visualization.* ACM. 373–382.

Marge, M., Banerjee, S., & Rudnicky, A. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. ICASSP, March.

Mason, W. A., & Watts, D. J. (2009). Financial incentives and the “performance of crowds.” *Proceedings of the Human Computation Workshop.* Paris: ACM.

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

Reips, U. (2002). Standards for Internet-Based Experimenting. *Experimental Psychology* (formerly "*Zeitschrift für Experimentelle Psychologie*"), 49(4), 243-256.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10 (pp. 2863–2872). New York, NY, USA: ACM.

Sala, M., Partridge, K., Jacobson, L., & Begole, J. (2007). An exploration into activity-informed physical advertising using pest. *Pervasive Computing*, 73–90.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 254-263. Honolulu, Hawaii: Association for Computational Linguistics.

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with Amazon Mechanical Turk. In Computer Vision and Pattern Recognition Workshops, 2008.

Stolee, K. T., & Elbaum, S. (2010). Exploring the use of crowdsourcing to support empirical studies in software engineering. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. 1–4.