How to Filter out Random Clickers in a Crowdsourcing-Based Study?

Sung-Hee Kim
School of Industrial Engineering
Purdue University
kim731@purdue.edu

Hyokun Yun
Department of Statistics
Purdue University
yun3@purdue.edu

Ji Soo Yi
School of Industrial Engineering
Purdue University
yij@purdue.edu

ABSTRACT

Crowdsourcing-based user studies have become increasingly popular in information visualization (InfoVis) and visual analytics (VA). However, it is still unclear how to deal with some undesired crowdsourcing workers, especially those who submit random responses simply to gain wages (random clickers, henceforth). In order to mitigate the impacts of random clickers, several studies simply exclude outliers, but this approach has a potential risk of losing data from participants whose performances are extreme even though they participated faithfully. In this paper, we used the randomness of multiple submissions from a crowdsourcing worker as a metric to determine whether or not a worker is a random clicker. Thus, we could reliably filter out random clickers and found that resulting data from crowdsourcing-based user studies were comparable with those of a controlled lab study. We also tested three representative reward schemes (piece-rate, quota, and punishment schemes) with four different levels of compensations ($0.00, $0.20, $1.00, and $4.00) on a crowdsourcing platform with a total of 1,500 crowdsourcing workers to investigate the influences that different payment conditions have on the number of random clickers. The results show that higher compensations decrease the proportion of random clickers, but such increase in participation quality cannot justify the associated additional costs.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology*

General Terms

Experimentation, information visualization.

Keywords

Crowdsouring-based user study, Amazon Mechanical Turk.

1. INTRODUCTION

The "crowdsourcing" approach [10] was originally adopted to accomplish difficult-to-be-automatized tasks (e.g., counting the number of persons in a photo) by harnessing human workers through an open call via online. It also has been adopted in various research disciplines, including information visualization (InfoVis) and visual analytics (VA), to conduct studies quickly and economically. The crowdsourcing approach has several advantages [17, 24]: First, one can efficiently recruit a large, diverse group of participants thanks to the active crowdsourcing platform (one can easily recruit 100 participants within 24 hours). Second, the cost for conducting a study through this approach is low. The average hourly wage of workers at Amazon Mechanical Turk (AMT), a crowdsourcing platform, is just under $2.00 [24]. Several studies (e.g., [8, 9, 22]) also reported that they successfully replicated the results of prior laboratory studies using crowdsourcing.

However, crowdsourcing also has limitations: one of the most serious issues is crowdsourcing workers often game an experiment system, such as randomly clicking through questions simply to earn monetary compensation [7]. Such behavioral trends reduce the validity of the collected data so researchers have investigated different methods to motivate workers and/or screen out spammers [6, 13, 27]. However, such approaches are usually not universally applicable in all types of studies, and some can damage the validity of a study by eliminating some outliers in an arbitrary manner [8, 29].

Thus, the goal of this paper is to propose a new approach to identify crowdsourcing workers who submit random responses, who we call "random clickers," and test the effectiveness of the approach. More specifically, we first investigated if the randomness metric can effectively filter out random clickers, so that the resulting data set is comparable to that from an equivalent controlled lab study. Second, we also investigated whether crowdsourcing workers are less likely to become random clickers if they receive higher compensation. We conducted an experiment to observe the influences of different compensation amounts ($0.00, $0.20, $1.00, and $4.00) and reward schemes (piece-rate, quota, and punishments) on the number of random clickers. These studies are the replication of the "SimulSort" study [11], which we previously conducted studies with in a controlled lab environment, using the crowdsourcing approach.

The contributions of this study are as follows:

- This study suggests the randomness metric, which can be used to discern random clickers.
- This study demonstrates the pseudo-linear relationships between the payment amounts and the likelihood of being random clickers.
- This study shows that payment schemes partially influence the likelihood of being random clickers.
- This study suggests guidelines for proper payments schemes in a crowdsourcing-based study in InfoVis and VA.

2. BACKGROUND

2.1 Crowdsourcing Quality Management

As a convenient and efficient recruiting platform, crowdsourcing has been quickly adopted as an alternative approach to conduct human subject studies. Recent studies [8, 9, 22] used the crowdsourcing approach to replicate previous controlled lab studies, to show that the results from crowdsourcing-based studies are comparable to those from controlled laboratory studies. However, researchers were still suspicious about the validity of crowdsourcing approach because crowdsourced research participants often game the system to simply gain monetary compensation without paying sufficient attention to the experimental tasks. The problem arise from a lack of control over the environment [4, 19, 28], difficulties in screening for well-balanced participants based on the demographical data [15, 26, 28], and uncertain data quality due to the low payment and anonymity of the Internet [18, 21, 22].

Thus, the following question is often asked, "how do we conduct a human subject study using the crowdsourcing approach and still obtain reliable results?" Generally, the following two approaches have been suggested: the first approach is to motivate the participants in the first place to perform in a sincere way. Obviously, monetary compensation is one of the main factors participating in crowdsourcing work, while in some cases it is actually a primary source of income for participants [22]. There have been attempts to change the compensation amount to nudge better performance, however the results of such attempts are obscure. More details will be discussed in Section 2.2. Another way to provoke intrinsic motivation is to design the task to be interesting. The enjoyment factor has been shown to be more powerful over the monetary values in some cases [21]. However, this is not always applicable, as it is hard to adjust the tasks to be entertaining in some experiments.

No matter how carefully the experiment is designed, experimenters are always exposed to the risk of having poor-quality data due to spammers. It is known that only about 60% of the collected data pass the qualification [7]. Therefore filtering out low quality work is important for quality management of the collected data. Researchers have proposed several ways of filtering such as inserting dummy questions to detect spammers [7], comparing the data to a gold standard [6], using post-hoc work assessment among workers [27], or creating algorithms to calculate the workers' quality [13]. However, in some cases inserting extra tasks could obstruct the experiment itself and gold standards may not exist. Therefore, prior researchers suggest screening based on the time spent [17], or using extreme values of the performance which they are trying to measure [8, 29].

Rzeszotarski and Kittur have pointed out that researchers should use *how* the workers work to detect low quality data collected [25]. They introduced a behavioral trace recording of all of the mouse movements and key presses. However, interpreting the logs are subjective and dependent on the task context. Important as it is, coming up with a proper metric that reflects the faithful attitude is still challenging [3].

2.2 Financial Rewards

We also consider the monetary factor, which is the primary reason for workers to assist on a crowdsourcing platform. Our approach concentrates on convincing the workers to partially mimic the "seriousness" of lab participants to obtain reliable results. From an economic perspective, the workers' goal is to maximize profit, while researchers want to guarantee quality. Therefore, we focus on two financial factors, the payment amount and the scheme the reward is designed to be given.

2.2.1 Payment Amount

Although the crowdsourcing platform has been widely used, there are not many empirical studies about the monetary effect on the participants. One of the initial studies conducted by Mason and Suri concluded that increasing workers' compensation does not increase their performance, only their quantity of work [20]. This was surprising due to the fact that the crowdsourcing platform used, AMT, is an online labor market where compensation is the main motive for participating. However, we should carefully interpret the results considering several factors. In Mason and Suri's experiment, the task was sorting images and solving word puzzles, and the incentives tested ranged from \$0.01 to \$0.10 which were given regardless of the workers' performance. In other words, the quality of the data submitted would not affect their earnings. It is noted that the amount only affected how they perceived the importance of the task; if the amount was higher, workers valued it more. Eventually, the higher compensation only attracted the participants to finish more tasks for higher earnings rather than promote better performance. Therefore, in order to make the amount have an impact on the work quality, a reward scheme that reflects the performance of the submitted work could be more effective.

2.2.2 Payment Scheme

Another speculation is that the incentive scheme was not effective enough to change the participants' behavior. Bonner et al. [5] suggested that in order to make the rewards effective, the proper amount should consider the task type based on the complexity. A wider range should be investigated, especially where our decision making task falls into the highest complexity "judgment and choice" task category. However, just raising the amount of money may not prevent all the "random-clicking" behavior; in fact, it may attract more participants to game the system. This is why we need to consider the following three financial reward schemes:

Piece-rate. Implementing piece-rate on crowdsourcing involves paying the workers for each task, which grows as they complete more tasks. Paying based on the number of tasks workers have finished is the standard practice on crowdsourcing platforms. However, the marginal benefit could be perceived smaller because the payment for one unit of task is rather small.

Quota. With a quota scheme, participants are paid a large amount only if they complete all the tasks, which motivates them to set a goal. It attempts to motivate people by increasing the perceived marginal benefits. The piece-rate versus quota payment has been investigated widely in economic discipline, and quota scheme usually out-performs piece-rate. Bonner et al. reviewed over 131 experiments, and concluded that quota schemes are the best reward schemes for complex tasks [5]. This was also found in some crowdsourcing studies involving the word puzzle task [21]. However, since implementing a quota scheme on the platform is against the norm and against what workers are accustomed to, we question whether it would be effective on the crowdsourcing platform.

Punishment. The last way to minimize the workers from gaming the system is to adopt a punishment scheme. Such schemes either do not pay workers or even deduct payment for poor performance. For a content analysis task on the crowdsourcing platform, a "punish agreement", where the researchers disclose that they will deduct 10% from the bonus when the work does not agree with the majority, was the most effective form of payment [27]. We believe that the risk of possibly losing money will prevent a considerable amount of random-clicking workers. Therefore, we added a penalizing scheme to pay a bonus only if the workers exceeded a certain level of performance, the criteria for which is based on the results from a controlled lab study. It could be argued that we are manipulating the workers to perform better. However, in reality, if workers were using this interface for their own need they would take it more seriously and pay more attention. Thus, we believe this constraint is not just artificial, but reflects reality.

2.3 SimsulSort

The complex task we use for the experiment is testing one of the InfoVis systems for decision making, called "SimulSort." SimulSort was proposed as an interactive table that sorts multiple attributes at the same time, supporting compensatory decision tasks [12]. In SimulSort, an item is not presented in a single row. Instead, SimulSort visualizes a cell at a position where the vertical position means the rank of the item in the corresponding attribute (Figure 1). The higher position a cell occupies, 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 overall values of an item in multiple attributes. The task given to workers in our study was to select the highest value items (sum of values across attributes). We added an incentive by giving participants higher payment as they select items more accurately. More details about SimulSort and its experimental design can be found in previous publications [11, 12].

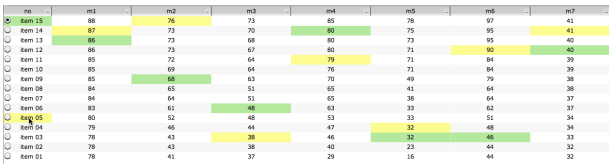

Figure 1: A screenshot of the SimulSort interface: comparison of the two items can be done by observing the vertical positions of the highlighted cells. Each color highlighted with green and yellow corresponds to one item.

3. METHODS

We replicated our previous controlled laboratory experiment [11] on one of the representative crowdsourcing platforms, Amazon Mechanical Turk (AMT). Participants were informed to complete a multi-attribute object selection task with SimulSort. We tested different reward amounts in different schemes to see how the performance was affected.

3.1 Participants

A total of 1,500 participants were recruited through AMT, 150 participants for each of the 10 conditions (see details in Section 3.4). Participants could only participate once for any of the conditions, as controlled by our system.

3.2 Procedures

After recruited from AMT, all the participants were redirected to our website for the experiment with an unique ID and password. Each participant was asked to complete 20 trials of tasks, where each trial had a 3-minute time limit. Instructions were given about the tool and the experiment procedure. The reward scheme and the maximum amount they could earn through the experiment were clearly mentioned. Before the 20 trials, a pilot trial was given to use the interface without time limit. After each trial, a dialog box with the result of the final selection rank and the amount they earned was shown. Participants were able to quit during the task and receive rewards commensurate with their progress. However, only the data from the participants who finished all 20 trials were used for analysis.

3.3 Tasks

The task was to select an item with the highest value out of 15 items with 7 attributes. The value of an item was calculated by the sum of its normalized attribute values as in the following equation 1: utility function $value_m$ is the mth item's value; T_{mn} is the number in the nth column (attribute) of the mth row (item) in data set T.

More specifically, the value was calculated as follows:

$$value_m = \sum_{n=0}^{7} \frac{T_{mn} - \min T_{\cdot n}}{\max T_{\cdot n} - \min T_{\cdot n}} \quad (1)$$

where $value_m$ is the value of the m^{th} item, T is the whole data set (15×7), and T_{mn} is the value of m^{th} item and n^{th} attribute.

Basically, $value_m$ is the summation of attribute values, each

of which is normalized within each column. Although this equation looks arbitrary, it was designed to avoid various types of gaming and resemble a real life decision-making, in which a person considers multiple attributes equally.

3.4 Experiment Design

We had a total of 10 conditions, with 3 payment schemes (i.e., piece-rate, quota, and punishment) associated with 3 reward amount levels (i.e., $0.20, $1.00, and $4.00) and no bonus rewards ($0.00).

Reward Amount. $0.10 was the base compensation for participating. The additional bonus reward was based on the task performance which was measured by the utility of the final selection. This was paid through the bonus payment service on AMT. We defined the range of reward amount based on the previous literature. The AMT system calculates the hourly compensation rate for each HIT which ranged up to approximately $8.00, while the average is $1.97. To select the proper amount for the experiment, the pilot study first paid from $0.00 to $8.00 at the completion of 20 trials. This also covers up to the federal minimum wage ($7.25/hour as of April 2, 2011). Based on our pilot study, we found out that the average time spent was approximately 30 minutes. Therefore for the actual experiment, we decided to pay a maximum of $4.00 to meet the federal minimum wage. The final amount level we tested for the bonus amount was, $0.00, $0.20, $1.00, and $4.00.

Piece-rate. Earnings for each round would add up to the total amount for the whole task. If the total maximum earning was $1.00, it would be divided by 20 rounds, so participants would earn $0.05 for each round.

Quota. We followed the same quota scheme design that was used in the previous controlled lab experiment. After finishing 20 rounds, only randomly selected three rounds were paid. For example, for the $1.00 bonus amount condition, each trial would have maximum earnings roughly $0.33 where the piece-rate scheme would have $0.05. This was to increase the perceived value of each round, so that the participants would be more motivated.

Punishment. This payment scheme includes a punishment based on the final selection for each round. The participants were paid for the trials only when their final choice was within the top 50%. In other words, to earn payment they should have selected an item ranked between the first and the seventh out of fifteen items.

3.5 Implementation

We implemented our experimental website using Ruby on Rails (`http://rubyonrails.org/`), Flex (`http://www.adobe.com/products/flex`), and Flare (`http://flare.prefuse.org`).

4. RESULTS

4.1 Presence of Random Clickers

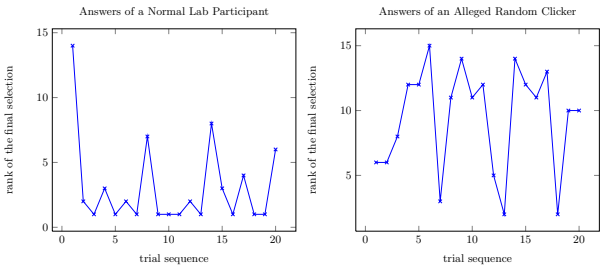

Figure 2: Example of answers of a lab participant (left) and an alleged random clicker (right). Rank 1 represents the highest rank while 15 is the lowest. Note that while the lab participant shows consistent performance, the ranks of items chosen by alleged random clicker fluctuate much more rapidly.

To make the notion of the random clicker more explicit, we first define the *rank* of the item to be the rank of the value of the item i:

$$rank_i = |\{i' : value_{i'} \geq value_i\}|, \quad (2)$$

where $|\cdot|$ denotes the number of elements in the set. Namely, $rank_i = 3$ implies that the the item i is the third best choice among 15 items.

The order of items displayed to the participant is designed to be well-permuted. Therefore, if the participant is not paying attention to the problem and choosing arbitrary items as answers (e.g., always selecting the item in the first row) then the ranks of the items chosen by the worker will be randomly distributed.

The presence of such random clickers in AMT-recruited participants is evident even with cursory visual inspection of data. Figure 2 shows an example of the answers of two different participants, one from a controlled lab study and one from crowdsourcing-based study. It is clear that the former shows consistent performance of selected items clustered on the top ranked items (Rank 1 represents the highest rank), while the latter exhibits very random behavior and his/her chosen items are widely scattered.

4.2 Threshold for Random Clickers

Proper filtering mechanisms have always been an interest in data collection on crowdsourcing platform [2, 3, 14]. If we take all of the data collected into consideration, it is obvious that we will not see similar results as seen from a controlled lab study. This was shown in one of the initial replication experiments from the authors [16]. We need an effective way to remove random clickers in analyzing the collected data. After removing such random data, one may question if the quality of the remaining data will be comparable with the data from the lab study. Using standard statistical testing, we propose a measure which quantifies the evidence of each participant being a random clicker. By filtering out participants according to this metric, we claim that with proper financial reward design, one can achieve the data quality that is comparable to lab experiments.

In order to screen out participants, we need to determine whether a certain participant is a random clicker or not.

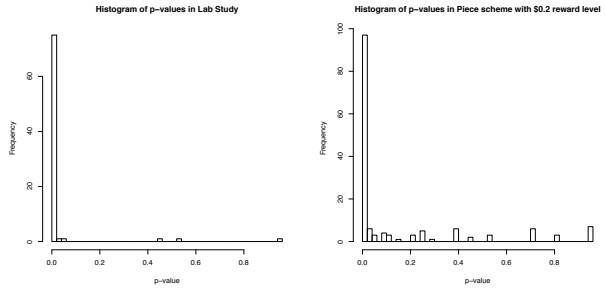

Figure 3: Histogram of p-values of participants in (left) lab study and (right) crowdsource data (piece scheme with $0.20 reward level was used), when each participant was tested by Pearson's χ^2 test.

This can be done by applying Pearson's χ^2 test [1]. For each participant, we test the null-hypothesis that he/she is a random clicker and thus his/her responses are uniformly distributed. As a result, we get a p-value for each user: a low p-value indicates that there is little evidence that the participant is a random clicker. As in the left of Figure 3, in lab study, most of p-values are concentrated towards the left around zero. This shows that it is very unlikely for a non-random clicker to get high p-value in χ^2 test. On the other hand, if the participant is a random clicker, then the p-value is uniformly distributed from 0 to 1. In the right figure of Figure 3, one can see that there are many data points uniformly distributed along the x axis. Therefore, by filtering out participants who have high p-values, one can achieve the data quality level which is comparable to lab studies.

A lower threshold implies that more participants are filtered: when the threshold is 0, no participant is included, while a threshold of 1 means that every participant is included. How we apply thresholds for filtering can yield a controversial issue related to forcing in order to refine the data. However, we can see that such filtering is inevitable for crowdsourced-based collected data. As Rzeszotarski and Kittur pointed out, we used the randomness as one way to reflect the sincerity of the participants and applied filtering based on it [25]. Assuming that the lab study is the standard quality to achieve, we define the threshold where the histogram declines extremely. In this experiment, we selected the p-value of 0.02 as shown in the left histogram of Figure 3.

4.3 Influence of Payment Amounts and Schemes

Table 1 shows the results after removing the random clickers based on threshold 0.02. We ran a non-parametric Mann-Whitney-Wilcoxon test to evaluate the difference of each crowdsourcing data against the lab data. We found significant difference from no reward condition (p-value < 0.001), piece-rate \$4.00 (p-value < 0.001), quota \$4.00 (p-value $= 0.001$), and punishment \$4.00 (p-value < 0.001). All three cases with \$4.00 had higher median and mean values than the lab data. The following median and mean values for all of the conditions can be found in Table 1.

Another point of interest was the actual number that remained after filtering, which is critical to the overall cost. The last column N in Table 1 is the number of remaining participants after filtering with threshold 0.02. The proportion of random clickers for each condition is shown in Figure 4. All three schemes have a similar trend proportional to the amount paid, decreasing as the payment level gets higher. Within each payment level, the punishment scheme has a relatively lower proportion than the other two schemes.

Setting	Level($)	Median	Mean	p-value	N	Cost($)
lab	-	0.979	0.887	-	80	1,750
no reward	0.00	0.937	0.865	**0.000**	93	0
piece-rate	0.20	0.970	0.878	0.172	97	31
	1.00	0.996	0.897	0.287	102	147
	4.00	1.000	0.911	**0.000**	125	480
quota	0.20	0.978	0.886	0.711	98	31
	1.00	0.979	0.887	0.963	109	138
	4.00	1.000	0.906	**0.001**	119	504
punishment	0.20	0.960	0.878	**0.023**	101	30
	1.00	0.996	0.901	0.117	110	136
	4.00	1.000	0.904	**0.000**	125	480

Table 1: Table of the median and mean values after filtering based on the threshold. The crowdsourcing data was compared to the lab data applying Mann-Whitney-Wilcoxon test. Bold p-values indicate that they are not comparable with the lab data. N is the number of participants remained after filtering from a total of 150 participants. Cost is calculated for the total payment when collecting 100 non-random clickers.

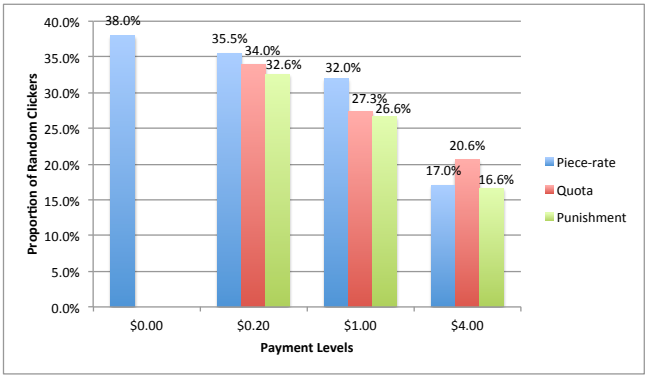

Figure 4: Proportion of random clickers calculated with different reward schemes and levels from Table 1. Note that in general the increase of the payment level decreases the proportion of random clickers.

5. DISCUSSION

We demonstrated that the consistency of performance calculated by *randomness* could be an implicit behavioral measure of the participants of crowdsourcing. After filtering the potential random clickers with a proper threshold, the quality of the crowdsourced data came out to be comparable to the lab study. We acknowledge the opinion that defining a threshold could be refining the data in an artificial way. However, in order to utilize the crowdsourcing data, such refinement is an inevitable process. After filtering, both \$0.20 and \$1.00 for piece-rate and quota scheme, as well as \$1.00 for punishment scheme, did not have a significant difference

with the lab study. In other words, this filtering enables researchers to use crowdsourcing data to substitute lab data. Interestingly, $4.00 for all schemes had significant difference with higher medians than the lab data. We believe that the amount of $4.00, which is quite a high payment on the crowdsourcing platform, attracted the participants to concentrate more on the task. Although, Mason and Suri mentioned that the amount paid does not affect the quality, we see that the mean and number that remained has a consistent trend of increasing when the payment gets higher [21].

As shown in Figure 4, the proportion of random clickers decreases as the level of financial reward gets higher. This provides a counter-example of the idea that the amount of payments does not increase the quality of the crowdsourcing work [20]. We believe that the findings in the previous literature are not always true when the payment is directly correlated with the performance and when additional payment that one can gain appears to be significant enough to warrant additional effort. The higher proportion of random clickers for $0.00 and $0.20 can be explained because of the low amount. Previous studies show that if the incentive scheme is not designed effectively, it will not yield a proper increase in performance [5]. Eventually the amount of $0.20 was less attractive to a larger portion of workers so that they easily exploited the system by randomly clicking.

While designing the study from the experimenter's perspective, maximizing the benefit with minimum cost is a point of consideration. Thus we must think of the total cost while deciding on the amount of the rewards. Although the $4.00 condition yields the highest decision quality, the experimenters can select the amount based on their experiment condition. We calculated the total cost for each condition in order to collect data from 100 non-random clickers. The total cost needed to collect data from 100 qualified participants is shown in the last column in Table 1. Even though $4.00 is the highest payment, it is still lower than the compensation given in the lab study. After selecting the reward amount, one should also consider the payment scheme. For example, within $0.20 and $1.00, selecting either quota scheme or penalty scheme will be effective, while piece-rate is less desirable as it has the highest portion of random clickers. Moreover, we believe that the fact that one can collect reliable data by spending only $31 is an attractive opportunity for experiments that need to conduct to wide range of participants with fast iterations for testing hypothesis [23].

At last, our experiment design had the advantage of being able to employ this consistency metric and make the financial value to be influential. The repeated measures design enabled us to calculate the randomness among each participant. Importantly, the amount of reward that the participants actually earned was directly linked to their performance. The performance and earning for each trial was shown after each round, in order to raise awareness and motivation. We believe that although there is a risk due to the uncertainty of the platform, well-designed experiments can control these factors to be effective.

6. CONCLUSIONS

One of the main questions we asked in this paper is whether crowdsourcing is a viable option for conducting user studies comparable to controlled lab studies, especially with an experiment that deals with complicated tasks such as decision making using visual interfaces. This question was raised to fill the gap of previous studies, many of which focused more on relatively short and simple perception oriented tasks, such as transcription and object recognition from images. Our answer to this question is "yes, but it is challenging." The main challenge appears to come from a lack of control over the environment, so crowdsourcing workers behave rationally to maximize their profit in any given time. If there is no marginal financial gain attractive enough to exert extra effort, it is more rational in terms of maximizing the return on investment to finish a task as quickly as possible and risk some penalties. What we demonstrated in this experiment is that there is a threshold of payment level that might change the workers' strategy of profit maximization, from increasing the number of tasks done in a given time to increasing the quality of the work.

Although we see that the crowdsourcing platform can handle usability studies with proper financial compensations, the overall quality of the data is not guaranteed to be as high as those from the lab study. Additional screening process is needed. The proportion of the random clicker in the lab study is very low, with only 1.5% from 80 participants. Considering the sincerity of the participants, there is some loss we need to expect in crowdsourcing. If we compare the cost and time for collecting enormous data, we believe it is still worthwhile. To run the experiment with 1,500 participants in a lab study, we would need approximately $30,000 and a longer period to schedule and conduct all of the studies. If we compare the cost and time for collecting enormous amount of data, we believe crowdsourcing approach is worth while to pursue, to fulfill the needs of user studies with diverse participation and fast iteration.

However, there are several limitations of our study. The threshold to filter random clickers is critical, yet it was selected with an arbitrary number. Other statistical approached such as Bayesian statistics could handle the selecting the threshold in a better way. More studies should be done such as 1) comparing the effectiveness of incentives and other fraud prevention methods, 2) comparing randomness and other filtering methods, and 3) testing different visualizations to generalize the findings.

References

[1] A. Agresti. *Categorical Data Analysis.* John Wiley & Sons, Inc., 2002.

[2] C. Akkaya, A. Conrad, J. Wiebe, and R. Mihalcea. 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*, pages 195–203, 2010.

[3] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. In *ACM SIGIR Forum*, volume 42, pages 9–15, 2008.

[4] M. Bloodgood and C. Callison-Burch. 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, pages 208–211. Association for Computational Linguistics, 2010.

[5] S. E. Bonner, R. Hastie, G. Sprinkle, and S. M. Young. A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research*, pages 19–64, Jan. 2000.

[6] C. Callison-Burch. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295, 2009.

[7] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 2399–2402, Atlanta, Georgia, USA, 2010. ACM.

[8] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 203–212, 2010.

[9] J. J. Horton, D. G. Rand, and R. J. Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425, 2011.

[10] J. Howe. Crowdsourcing: A definition. *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business*, June 2006.

[11] I. Hur, S.-H. Kim, A. Samak, and J. S. Yi. A comparative study of three sorting techniques in performing cognitive tasks on a tabular representation. *International Journal of Human-Computer Interaction*, 2012. doi:10.1080/10447318.2012.713802.

[12] I. Hur and J. S. Yi. SimulSort: multivariate data exploration through an enhanced sorting technique. In *Human-Computer Interaction. Novel Interaction Methods and Techniques*, volume 5611, pages 684–693. Springer Berlin Heidelberg, 2009.

[13] P. G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.

[14] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67, 2010.

[15] P. Kelley. Conducting usable privacy & security studies with amazon's mechanical turk. In *Symposium on Usable Privacy and Security (SOUPS)*, 2010.

[16] S.-H. Kim, S. Li, B. c. Kwon, and J. S. Yi. Investigating the efficacy of crowdsourcing on evaluating visual decision supporting system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 55(1):1090–1094, Sept. 2011.

[17] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456, Florence, Italy, 2008. ACM.

[18] R. Kosara and C. Ziemkiewicz. Do mechanical turks dream of square pie charts? *BELIV*, 10:373–382, 2010.

[19] M. Marge, S. Banerjee, and A. Rudnicky. Using the amazon mechanical turk for transcription of spoken language. *ICASSP, March*, 2010.

[20] W. Mason and S. Suri. Conducting behavioral research on amazon's mechanical turk. *SSRN eLibrary*, Oct. 2010.

[21] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.*, 11(2):100–108, 2009.

[22] G. Paolacci, J. Chandler, and P. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5), 2010.

[23] D. Rand. The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 2011.

[24] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. 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*, pages 2863–2872, Atlanta, Georgia, USA, 2010. ACM.

[25] J. M. Rzeszotarski and A. Kittur. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. 2011.

[26] M. Sala, K. Partridge, L. Jacobson, and J. Begole. An exploration into activity-informed physical advertising using pest. *Pervasive Computing*, pages 73–90, 2007.

[27] A. D. Shaw, J. J. Horton, and D. L. Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 275–284, 2011.

[28] K. T. Stolee and S. Elbaum. 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*, pages 1–4, 2010.

[29] C. Ziemkiewicz and R. Kosara. Laws of attraction: From perceived forces to conceptual similarity. *Transactions on Visualization and Computer Graphics (Proceedings InfoVis)*, 16(6):1009–1016, 2010.