Crowdsourcing Performance Evaluations of User Interfaces

Steven Komarov, Katharina Reinecke, Krzysztof Z. Gajos
Intelligent Interactive Systems Group
Harvard School of Engineering and Applied Sciences
33 Oxford St., Cambridge, MA 02138, USA
{komarov, reinecke, kgajos}@seas.harvard.edu

ABSTRACT

Online labor markets, such as Amazon's Mechanical Turk (MTurk), provide an attractive platform for conducting human subjects experiments because the relative ease of recruitment, low cost, and a diverse pool of potential participants enable larger-scale experimentation and faster experimental revision cycle compared to lab-based settings. However, because the experimenter gives up the direct control over the participants' environments and behavior, concerns about the quality of the data collected in online settings are pervasive. In this paper, we investigate the feasibility of conducting online performance evaluations of user interfaces with anonymous, unsupervised, paid participants recruited via MTurk. We implemented three performance experiments to re-evaluate three previously well-studied user interface designs. We conducted each experiment both in lab and online with participants recruited via MTurk. The analysis of our results did not yield any evidence of significant or substantial differences in the data collected in the two settings: All statistically significant differences detected in lab were also present on MTurk and the effect sizes were similar. In addition, there were no significant differences between the two settings in the raw task completion times, error rates, consistency, or the rates of utilization of the novel interaction mechanisms introduced in the experiments. These results suggest that MTurk may be a productive setting for conducting performance evaluations of user interfaces providing a complementary approach to existing methodologies.

Author Keywords

Crowdsourcing; Mechanical Turk; User Interface Evaluation

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces-Evaluation/Methodology

INTRODUCTION

Online labor markets, such as Amazon's Mechanical Turk (MTurk), have emerged as an attractive platform for human subjects research. Researchers are drawn to MTurk because the relative ease of recruitment affords larger-scale experimentation (in terms of the number of conditions tested and the number of participants per condition), a faster experimental revision cycle, and potentially greater diversity of participants compared to what is typical for lab-based experiments in an academic setting [15, 24].

The downside of such remote experimentation is that the researchers give up the direct supervision of the participants' behavior and the control over the participants' environments. In lab-based settings, the direct contact with the experimenter motivates participants to perform as instructed and allows the experimenter to detect and correct any behaviors that might compromise the validity of the data. Because remote participants may lack the motivation to focus on the task, or may be more exposed to distraction than lab-based participants, concerns about the quality of the data collected in such settings are pervasive [11, 19, 23, 26].

A variety of interventions and filtering methods have been explored to either motivate participants to perform as instructed or to assess the reliability of the data once it has been collected. Such methods have been developed for experiments that measure visual perception [8, 12, 3], decision making [20, 27, 14], and subjective judgement [11, 19, 21].

Missing from the literature and practice are methods for remotely conducting performance-based evaluations of user interface innovations — such as evaluations of novel input methods, interaction techniques, or adaptive interfaces — where accurate measurements of task completion times are the primary measure of interest. Such experiments may be difficult to conduct in unsupervised settings for several reasons. First, poor data quality may be hard to detect: For example, while major outliers caused by a participant taking a phone call in the middle of the experiment are easy to spot, problems such as a systematically slow performance due to a participant watching TV while completing the experimental tasks may not be as easily identifiable. Second, the most popular quality control mechanisms used in paid crowdsourcing, such as gold standard tasks [1, 18], verifiable tasks [11], or checking for output agreement [13, 17], do not have obvious equivalents in this setting.

In this paper, we investigate the feasibility of conducting remote performance evaluations of user interfaces with anonymous, unsupervised, paid participants recruited via MTurk. We implemented three performance experiments to

CHI'13, April 27–May 2, 2013, Paris, France.

re-evaluate three previously well-studied designs: the Bubble Cursor [7], the Split Menus [25], and the Split Interfaces [5]. We conducted each experiment both in lab with participants recruited from the local community and online with participants recruited via MTurk. We focused our investigation on three questions: (1) Would the data collected in the two settings demonstrate the presence of the same statistically significant differences? (2) If so, would the effect sizes be comparable across the two settings? (3) Would the absolute values of the measurements obtained in the two settings be similar?

The analysis of our results did not yield any evidence of significant or substantial differences in the data collected in the two settings. All statistically significant differences detected between experiment conditions in lab were also present on MTurk and the effect sizes were very similar. In addition, there were no significant differences in the raw task completion times, error rates, consistency, or the rates of utilization of the novel interaction mechanisms introduced in the experiments.

These results suggest that MTurk may be a productive setting for conducting performance evaluations of user interfaces.

In summary, in this paper we make the following contributions:

1. We present results of three experiments conducted on MTurk and in lab. By showing that the results are closely matched, we build a case for the feasibility of MTurk as a tool for user interface research.
2. We synthesize a number of practical considerations that had substantial impact on the quality of the data we collected. These considerations included mechanisms for ensuring instruction comprehension, accounting for age- and device-related differences in performance, techniques for robust outlier detection, and implementation challenges.

PRIOR RESEARCH

In behavioral economics, there is mounting evidence that MTurk participants perform just like lab-based participants. Replications of a number of classic experiments including those related to judgement and decision-making [20], public goods [27], and others [9, 22], have all shown no differences between results collected on MTurk and those collected in lab.

Similarly, Heer and Bostock [8] presented compelling evidence that experiments related to graphical perception can be reliably conducted on MTurk. However, their work also identified several potential pitfalls for experimenters. For example, variations in hardware and software had significant impact on the results — keeping careful records of these factors and controlling for them in the analysis overcame the problem. Also, many participants demonstrated signs of not having understood the instructions, presumably because they hurried through that part of the experiment. Qualification tasks, which required Turkers to demonstrate proficiency with a concept or skill, proved to be an adequate mechanism for enforcing instruction comprehension. Novel research has already been published in graphics and information visualization based primarily on data collected on MTurk [3, 12].

In contrast, survey-based research and research collecting simple subjective judgements have been less successful in leveraging MTurk. A number of researchers reported problems with persistently low-quality data or outright malicious participant behavior [11, 19, 26]. These results are frequently attributed to two properties of surveys and similar instruments: the truthfulness of responses is hard to verify, and the effort required to provide truthful and accurate responses is substantially higher than just selecting random responses. This attracts “spammers,” who often find that they can earn the payoff easily with little chance of being caught.

To mitigate the data quality problems on MTurk, researchers have investigated several classes of approaches. Gold standard tasks [1, 18], instructional manipulation check [19], the Bayesian Truth Serum [21], or certain behavioral measures [23] have all been used to detect low quality data and unreliable participants. Excluding such data from analysis can substantially improve the reliability and the statistical power of the subsequent analyses [19]. An ideal solution, however, would prevent low quality data from being collected in the first place. Data quality can be improved, for example, if participants are forced to see the question for a certain “waiting period” [10]. However, attempts to manipulate participants’ intrinsic motivation have been generally unsuccessful in improving quality of the work [2, 26]. Financial incentives generally help with faster recruitment and may incentivize participants to do more work, but generally do not impact the quality of the work either [16, 26] (though there is some evidence that quality-related financial bonuses or punishments may help [26]).

So far, we are not aware of any systematic evaluation of MTurk as a platform for performance-based evaluations of user interface technologies. There are good reasons to believe that such tasks will fare well on MTurk: Turkers’ incentives to finish the task as quickly as possible align with the goals of the experimenters. However, the unsupervised and uncontrolled nature of the participants’ environments leaves open the possibility that social and environmental distractions will introduce unacceptable variance or a systematic bias into the data.

EXPERIMENTS

We begin by investigating whether experiments comparing user performance on two or more user interface variants yield the same results both in lab and on MTurk. Specifically, we ask two questions. First, are the conclusions of the statistical comparisons of the primary performance measures (task completion times and error rates) between experimental conditions the same in both lab and MTurk settings? Second, are the effect sizes comparable?

Methodology

To answer these questions, we conducted — both in lab and on MTurk — three experiments evaluating three previously investigated user interface designs. These designs were the *Bubble Cursor* [7], the *Split Menus* [25], and the *Split Interfaces* [6]. These experiments are illustrative of the types of evaluations that are the focus of this paper: the experimental tasks are largely mechanical (i.e., target acquisition with a mouse pointer) as opposed to cognitively demanding (e.g., solving a problem, generating a creative artifact). Also, each of the three experiments we chose required a different level of attention from the participants. The baseline condition in the Bubble Cursor experiment was comprised entirely of simple pointing tasks requiring minimal cognitive engagement. In contrast, Split Menus and Split Interfaces are adaptive techniques that operate by copying useful functionality to a more convenient location. To reap the performance benefits afforded by these adaptive mechanisms, the user has to consciously monitor the adaptive part of the user interface to determine if a useful adaptation has taken place. In prior studies, participants reported that scanning for useful adaptations required conscious effort and most participants missed useful adaptations at least part of the time [5, 6]. Because concerns about online participants' lack of attentiveness are common [19, 20], we included the Split Menus and the Split Interfaces as probes of the differences in cognitive engagement between lab- and MTurk-based participants.

We implemented all three experiments for web-based delivery. We used the same software and the same instructions in both the lab and MTurk settings. For lab experiments, we recruited between 10 and 14 subjects each. These numbers were similar to those used in the original studies and were sufficient to demonstrate the main significant differences. For experiments conducted on MTurk we recruited approximately 100 participants for each experiment taking advantage of the ease of recruitment and the negligible marginal effort required to include additional participants.

Lastly, we sought to establish a clear criterion for identifying and excluding extreme outliers. A common approach in HCI research is to exclude results that are more than two standard deviations from the mean. This approach is not robust in the presence of very extreme outliers that are different from the rest of the data by orders of magnitude because such outliers introduce large errors to the estimates of the mean and the standard deviation. Unfortunately, such outliers, while rare in general, can be expected on MTurk and did, indeed, occur in our data. For example, one participant spent over 2 minutes on a trivial target selection task that he accomplished previously in 1.5 seconds — presumably the participant got distracted by an external event in the middle of the experiment.

To guard against such extreme outliers, we chose an outlier removal procedure that relies on a more robust statistic, the inter-quartile range (IQR), which is defined as the difference between the third and first quartiles. With this procedure, an extreme outlier is one that is more than $3 \times IQR$ higher than the third quartile, or one that is more than $3 \times IQR$ lower than the first quartile [4]. For normally distributed data, this procedure would remove less than 0.00023% of the data compared to 4.6% removed by the more typical mean $\pm$ 2 standard deviations approach. Thus, it targets the most extreme outliers without reducing legitimate diversity of the data.

We performed outlier detection based on two measurements: log-transformed per-participant mean selection time and log-transformed per-participant maximum selection time. A participant was removed from analysis if he or she was flagged as an outlier according to either measurement.

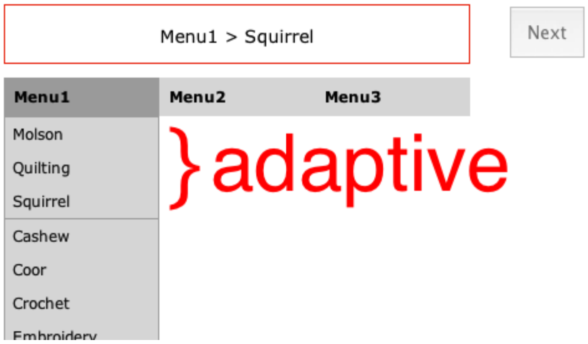


Figure 1: Split Menus. The three most recently selected items are copied in the adaptive top part of the menu.

This proved to be both a robust and an acceptably conservative approach. While 30% exclusion rates are typical for MTurk, and some researchers discarded data from up to half of their participants [8], our procedure resulted in the exclusion of between 0% and 6% of MTurk participants in our three experiments.

Experiment 1: Adaptive Split Menus

In Split Menus, a small number of menu items that are predicted to be immediately useful to the user are copied[1] to a clearly designated adaptive area at the top of the menu (Figure 1). The items copied to the adaptive area represent the system's best prediction for the user's next selection. The choice of a predictive algorithm varies across implementations, but the use of the the most recently used (MRU) and the most frequently used (MFU) algorithms is common.

Whether the Split Menus improve performance depends on the accuracy of the predictive algorithm and on the willingness of the user to pay attention to the adaptive portion of the menu.

Tasks and Procedures

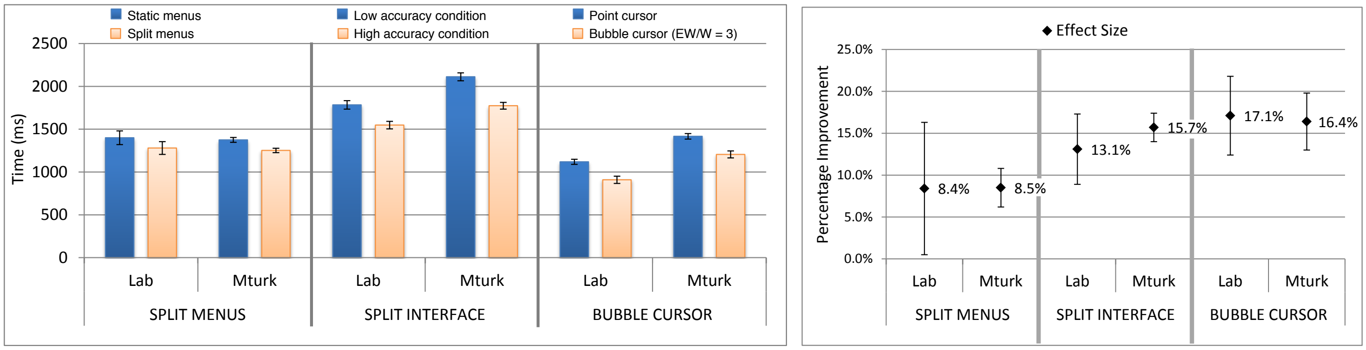
In our experiment we used a menu with three categories (Menu1, Menu2, Menu3) each containing 16 items, ordered alphabetically. The adaptive portion at the top of each category contained the three most recently used items, which were initialized randomly at the beginning of each block. For each category we generated a random sequence of 60 selections, constrained such that for 79% of the selections the goal item was in the adaptive portion. The three sequences of 60 selections were then randomly shuffled to obtain the final a sequence of 180 selections.

As the control condition, we used a typical static menu design, which differed from the split menu only in the lack of an adaptive portion at the top. At the beginning of the

[1] In the original design [25] items were *moved* rather than copied, but in all modern implementations of this concept items are copied.

Experiment	Participants				Task completion time (ms)				Errors			
Split Menus	Setting	rem	out	kept		Static	Split	ANOVA		Static	Split	Wilcoxon Signed-Rank
	Lab	0	0	14		1400	1280	F(1,13)=5.8, p=0.032		2.43	1.71	z=–0.14, p=0.887
	MTurk	6	1	89		1375	1252	F(1,88)=56, p<0.0001		2.45	2.22	z=–0.95, p=0.343
Split Interface	Setting	rem	out	kept		Low	High	ANOVA		Low	High	Wilcoxon Signed-Rank
	Lab	0	1	9		1784	1548	F(1,8)=42, p<0.0001		4	2.56	z=–1.38, p=0.169
	MTurk	6	0	86		2112	1774	F(1,85)=285, p<0.0001		3.87	3.55	z=–1.99, p=0.046
Bubble Cursor	Setting	rem	out	kept	point	b1	b3	ANOVA	point	b1	b3	Friedman
	Lab	0	1	12	1119	1122	909	F(2,22)=65, p<0.0001	7.58	2.33	1.08	$\chi^2(2)$=17, p<0.0001
	MTurk	9	6	108	1417	1410	1205	F(2,214)=110, p<0.0001	8.34	3.4	1.77	$\chi^2(2)$=108, p<0.0001

Table 1: All experiments: main effects. The "rem" column shows the number of participants removed due to a self-reported faulty device or medical condition. The "out" column shows the number of participants removed as extreme outliers.



(a) Mean selection times. Error bars represent SEM.

(b) Effect sizes. Error bars represent 95% CI.

Figure 2: Task completion times and effect sizes in all three experiments.

experiment, each participant completed a 30 selection split-menu practice block. Next, they performed four experimental blocks: two 90-selection split-menu blocks and two 90-selection static-menu blocks, alternating (the selection of the starting condition was counterbalanced across participants). At the beginning of each trial, participants had to click the "Next" button for the prompt to display the next target item. The experimental UI is shown in Figure 1.

Design and Analysis

We used a within subjects design with one factor (the menu design) with two levels (static or split). We analyzed two measures: trial completion time (measured from clicking on the category header to the correct item selection within the category), and number of errors per block. We recorded an error if a participant clicked on an incorrect item in a menu or clicked on an incorrect category label. However, the latter proved unnecessary as practically all errors were item errors. We log-transformed the timing data to account for the skewed distributions found in such data. We analyzed the timing data using repeated measures ANOVA and we used the Wilcoxon signed-rank test for the error analysis.

Participants

We recruited 14 participants (10 male, 4 female) aged 18–35 (M=26) from our university to participate in the lab version of the study. We recruited 96 MTurk participants (US-based, 95% approval rate, minimum of 500 completed HITs, 49 male, 47 female) aged 18–65 (M=30) for the online variant of the study.

Results

Adjustments of Data

There were 6 participants on MTurk who reported having a medical condition or a faulty hardware device that might have interfered with their work, and we discarded their data without further inspection. Among the remaining 90 participants, 1 was classified as an extreme outlier and was removed from analysis. In lab, there were no extreme outliers.

Main Effects

The results are summarized in Table 1 and are visualized in Figure 2(a). In both settings (lab and MTurk) we observed the main effect of the menu type (split vs. static) on task completion time. In both settings, participants were significantly faster with the adaptive split menus than with the static baseline design. All participants committed fewer errors in the split condition, however the differences were not significant.

There were no substantial or significant differences in the magnitude of the effect size between the two settings (Figure 2(b)).

Experiment 2: Adaptive Split Interface

Split Interfaces are a generalization of Split Menus. In Split Interfaces, the items predicted to be most immediately useful to the user are copied to specially designated part of the user interface [5]. Therefore, just as in Split Menus, when the system correctly places the desired item in the adaptive part of the interface the user has the choice to either use the adaptive copy or to access the item at its usual location in the static part of the interface.

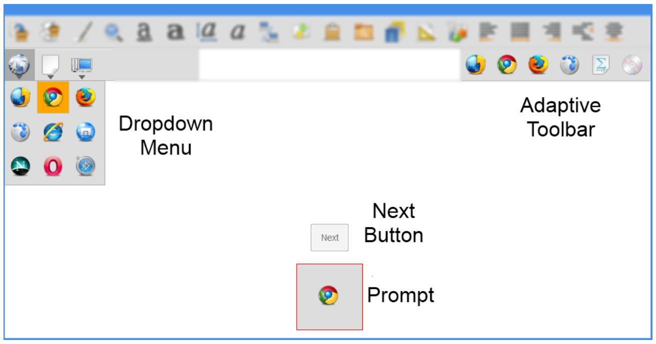


Figure 3: Split Interface. The 6 most recently used items (2 per category) are copied to the adaptive toolbar.

We reproduced an experiment, which demonstrated the impact of the accuracy of the predictive algorithm on users' performance [6]. That is, the experiment compared two otherwise identical variants of a Split Interface (illustrated in Figure 3), which differed only in their behavior: in one condition the desired item was present in the adaptive part of the interface with a 50% probability and in the other with a 70% probability.

In our experiment, the adaptive part of the interface was placed at the opposite side of the screen from the static part of the interface used during the experiment. Therefore, to take advantage of the adaptation, users had to consciously direct their gaze to the adaptive part of the interface.

Tasks and Procedures

We used a design with three categories of items for users to select: browsers, file types, and computer devices. Each category had 9 items arranged in a 3-by-3 square-grid formation. Each category was represented as a button on a toolbar. Clicking this button opened up a drop down menu containing all the items in that category. The adaptive toolbar contained the two most recently selected items per category for a total of six items. At the beginning of each trial, a copy of the target item, used as the prompt, was displayed in the middle of the screen. Also located there was a "Next" button used to proceed to the next selection. Thus, the typical selection trial consisted of clicking the "Next" button, looking at the prompt to learn the appearance of the target, and selecting the target from either the adaptive toolbar or from the drop down menu.

Each participant started with a practice block of 10 selections in which they had to use the adaptive toolbar at least once or otherwise were asked to repeat the block. This ensured that the participants understood the purpose of the adaptive toolbar. Next, they completed a second practice block consisting of 60 selections across all three categories such that in 60% of the selections the adaptive toolbar had a copy of the target. During this second practice block and in the main experiment, the use of the adaptive toolbar was left entirely to each participant's discretion. The main experiment consisted of two blocks with 60 selections each. One of the blocks had predictive accuracy of 70% of containing the target item (high accuracy condition) and the other had predictive accuracy of 50% (low accuracy condition). The ordering of high and low accuracy conditions were counter-balanced across participants. The same selection sequences were used across all participants.

Design and Analysis

We used a within-subjects design with one factor (the prediction accuracy) with two levels (low=50% and high=70%). We measured two variables: trial completion time (measured from clicking on the "Next" button to the selection of the target in either the regular menu or in the adaptive toolbar), and number of errors per block. There were two types of errors: when the participant selected a wrong category and when the participant selected a wrong item. We used the sum of the two in the analysis. We log-transformed the timing data to account for skewed distributions found in such data. Timing data were analyzed with a repeated measures ANOVA. Errors were analyzed with the Wilcoxon signed-rank test.

Participants

For the lab study we recruited 10 students from our university (6 male, 4 female) aged 21–34 (M=26.5). From MTurk we recruited 92 participants(US-based, 95% approval rate, minimum of 500 HITs, 54 male, 38 female) aged 18–63 (M=27).

Results

Adjustments of Data

There were 6 participant on MTurk whose data was discarded because they reported a medical condition or an unreliable pointing device that might have affected their performance. There were no extreme outliers on MTurk. In lab one participant was removed as an extreme outlier, because he was substantially slower than the rest of the participants for a reason we could not determine.

Main Effects

As before, the results are summarized in Table 1 and visualized in Figure 2(a). We observed a significant main effect of the predictive accuracy on task completion time both in lab and on MTurk. In both settings participants were faster in the high accuracy condition, and there was no evidence of a substantial difference in effect size (Figure 2(b)). In addition, participants in both settings made fewer errors in the high accuracy condition. On MTurk the difference was significant, while in lab it was not (Table 1).

Experiment 3: Bubble Cursor

The Bubble Cursor [7] is a pointing enhancement where the size of the cursor's hot spot changes dynamically to always overlap the closest target. The Bubble Cursor will completely contain the closest target (Figure 4(b)), unless this would cause it to intersect the second closest target. In that case, the size of the cursor is reduced and the primary target is decorated with an outline as shown in Figure 4(c). The Bubble Cursor takes advantage of the sparsity of the targets to increase their effective sizes. The more sparsely the targets are distributed, the greater the potential benefits conferred by this design compared to the traditional point cursor. We reproduced a study comparing the Bubble Cursor (at two different target density settings) to the point cursor.

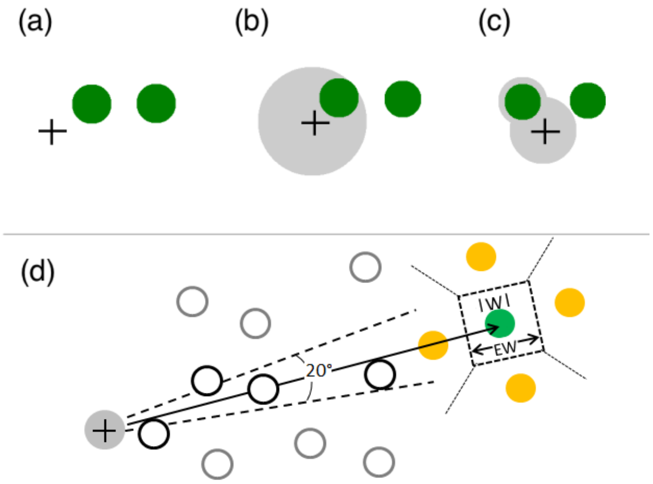


Figure 4: Point and Bubble Cursors. (a) Point (crosshair) cursor. (b) and (c) The size of the hotspot of the bubble cursor changes dynamically to intersect with the closest target while avoiding any other target. (d) The *effective width (EW)* of the target can be manipulated by changing the proximity of the closest four distracter targets (shown in yellow) to the goal target (shown in green). This figure based on [7].

Tasks and Procedures

There were five independent variables in the original Bubble Cursor experiment. Cursor type $CT = \{Point, Bubble\}$; amplitude A (i.e., the distance to the target); target width W; the ration of effective width to width EW/W, where effective width is the width of the target W plus all surrounding pixels that are closer to it than to any other target (as illustrated in Figure 4(d)); and distracter density D, which, as also illustrated in Figure 4(d), is the density of the targets in the 20 degree slice between the starting point and the target.

In our replication of the experiment, we varied only the cursor type CT and the EW/W ratio, while fixing the other variables: W=12px, D=0.5, A=Uniform(200px,400px).

Each participant completed a practice block (CT=Bubble, EW/W=3) consisting of 30 target selections. Participants were required to use the "bubble" capability of the Bubble Cursor at least once during the practice block (i.e., they had to acquire a target such that the center of the cursor was outside the target, but the target was captured by the "bubble") or else they were forced to repeat the practice block. This ensured that all participants understood the special capability of the Bubble Cursor. In all other blocks, the use of the "bubble" mechanism was used to each participant's discretion.

The practice was followed by three 80-selection blocks ordered randomly for each participant: (baseline condition with CT=Point), (CT=Bubble, EW/W=3), and (CT=Bubble, EW/W = 1.33). For convenience, we will refer to these conditions as $point$, $b3$, $b1$, respectively. As in the original version of the experiment, the target was rendered as green disc, the distracters as grey discs, and the bubble cursor as a semi-transparent grey disc. Both the active target and the distracters were filled with red when the cursor captured them.

Design and Analysis

We used a within subjects design with one factor (cursor type) with three levels (point, b1, b3). We measured two variables: trial completion time (measured from clicking on the previous target to clicking on the current target), and the number of errors per block (errors were defined as clicks that did not capture the target). We log-transformed the timing data to account for the skewed distributions found in such data. We analyzed the timing data using repeated measures ANOVA. For the error data, we used the Friedman test.

Results

Participants

We recruited 13 participants from our university (10 male, 3 female) aged 21–53 (M=25), and 123 MTurk participants (US-based, 95% approval rate, minimum of 500 HITs), (65 male, 58 female) aged 18–68 (M=27.5).

Adjustments of Data

There were 9 MTurk participants who reported having a medical condition or an unreliable input device that might have interfered with their work, and we discarded their data without further inspection. From the remaining 114 participants 6 were excluded from the analysis as extreme outliers. In lab, one participant was excluded because he mistakenly used a touchpad instead of the provided mouse.

Main Effects

As before, the results are summarized in Table 1 and visualized in Figure 2(a). We observed the main effect of cursor type on task completion time in both the lab setting and on MTurk, and the pairwise comparison (Bonferroni adjusted) showed significant difference ($p < 0.0001$) between $point$ and $b3$, and $b1$ and $b3$ for both lab and MTurk participants (Figure 2(a)). In neither setting was the difference between $point$ and $b1$ significant. There was no significant difference in effect size between lab and MTurk (Figure 2(b)). All participants committed the most errors in the $point$ condition, and the fewest in the $b3$ condition. The effect of cursor type on error rate was significant in both the lab and the MTurk settings.

COMPARISON OF LAB VS. MTURK RESULTS

The results in the previous section suggest that the *relative* differences between experimental conditions are the same whether measured in lab or remotely with participants recruited via MTurk. In this section, we address the question of whether there are systematic differences in the *absolute* magnitudes of the measurements collected in lab and on MTurk. Is either population systematically faster, more accurate, or more consistent? We compare the populations of the lab and MTurk participants in terms of speed, error rates, consistency, and utilization of the novel interactive mechanism presented in each experiment. We also characterize the demographics of the two populations. Finally, we investigate the reproducibility of the MTurk-based results.

Methodology

To compare the lab and MTurk populations, we captured the following measures:

- **Mean task completion times.**
- **Consistency.** We used the per-participant standard deviation in task completion times as a measure of how consistently they performed throughout the experiment.
- **Error rates.**
- **Utilization of the novel interactive mechanisms.** Each of our experiments included a novel interaction mechanism (adaptive area at the top of the Split Menus, adaptive toolbar in the Split Interface, and the ability to acquire the target without placing the pointer directly over it in Bubble Cursor). In all experiments, the use of these novel interaction mechanisms was designed to improve performance, but was optional and required a small amount of cognitive effort to use. We define utilization as the fraction of times when the user used the novel interaction mechanisms when one was available.
- **Fraction of extreme outliers.** This measure captures the fraction of participants who were excluded from analysis because they were classified as extreme outliers.

As before, we log-transformed the timing data before analysis. We used ANOVA to analyze timing, consistency, and utilization data. We used the (between-subjects) Wilcoxon rank-sum test to analyze error data.

The Bubble Cursor experiment posed additional challenges for the analysis, because computer performance had a significant effect on participant performance. The experiment was implemented using the HTML5 canvas element, whose performance depends on browser type, browser version, operating system, graphic acceleration, and CPU speed. Aware of these differences, we implemented an automatic performance check: Turkers whose browsers did not appear capable of redrawing the canvas in 30 ms or less were not allowed to proceed with the experiment. We determined this threshold as the slowest drawing speed at which the performance of Bubble Cursor appeared smooth and did not register any perceptible lag. For reference, the computer used in the lab study performed at 6 ms per frame (MSPF). The automatic performance check provided only an estimate of actual performance so we also logged the actual drawing performance during the experiment. In reality, the Turkers' computers performed in the range of 4 to 50 MSPF. Contrary to our expectations, even such short drawing times had a significant impact on performance: drawing time was positively correlated with selection time ($r = 0.267, n = 324, p < 0.0001$) and negatively correlated with errors ($r = -0.121, n = 324, p = 0.029$).

To account for different distributions of the background variables related to the participants' demographics and environment we used two redundant, parallel methods of analysis, both of which led to the same conclusions. In the first method, we included gender, input device, computer performance (bubble cursor only), and age as factors or covariates in the general linear model associated with the ANOVA. The advantage of this approach is that it leveraged all available data, but as a linear regression model it made two assumptions: (1) participant performance was linear with age and computer performance and (2) gender and device had a constant additive effect. In our second method we removed the need of the first assumption by matching the data in terms of age and computer performance using cutoff values. Results from both methods are shown in Table 2, in which all reported means (except for errors) are least-squares adjusted means computed based on the estimated parameters of the linear regression.

Results

Table 2 summarizes the results from this analysis.

Split Menu

Both analyses indicated that Turkers were slightly faster than lab participants, had lower individual standard deviations, had higher utilization rate, but made more errors. However, the differences were not significant. These differences were also relatively small: the differences in task completion times between the two populations were 3.5% in the analysis that included all data and 2.8% in the analysis with matched data. In comparison, the performance differences between the two experimental conditions (the static menus and the Split Menus) were larger than 8.4% (Figure 2(b) in the previous section).

Split Interface

In both analyses, Turkers were slower than lab participants, had higher standard deviations, made more errors, and had lower utilization rates. As before, the differences were not significant. The differences in task completion times between the two populations were 5.9% in the analysis that included all data and 6.3% in the analysis with matched data. In comparison, the performance differences between the two experimental conditions (as reported in previous section) were larger than 13.1%.

Bubble Cursor

In both analyses, Turkers were slower, had higher individual standard deviations, made more errors, and had lower utilization rates, but none of the differences were significant. The differences in task completion times between the two populations were 3.0% in the analysis that included all data and 4.9% in the analysis with matched data. In comparison, the performance differences between the two experimental conditions (as reported in previous section) were larger than 16.4%.

			Task completion time		Individual St. D.		Utilization		Errors					
			Mean	Significance	Mean	Significance	Mean	Significance	Mean	Sig.	Mean	Sig.	Mean	Sig.
split menus									Split		Static			
	all age	Lab	*1373*	F(1,200)=0.931,	*643*	F(1,200)=0.002,	*0.811*	F(1,98)=1.839,	1.71	z=0.571	2.43	z=0.737		
		MTurk	*1327*	p=0.336	*640*	p=0.963	*0.875*	p=0.178	2.22	p=0.568	2.45	p=0.461		
	age<35	Lab	1310	F(1,133)=0.057,	629	F(1,133)=0.401,	0.841	F(1,65)=0.033,	1.85	z=0.838	2.62	z=0.904		
		MTurk	1274	p=0.450	599	p=0.527	0.863	p=0.857	2.63	p=0.402	2.75	p=0.366		
split interface									High		Low			
	all age	Lab	*1846*	F(1,184)=2.129,	*791*	F(1,184)=2.231,	*0.959*	F(1,90)=0.731,	2.56	z=0.064	4	z=-0.781		
		MTurk	*1955*	p=0.146	*971*	p=0.137	*0.912*	p=0.382	3.55	p=0.949	3.87	p=0.435		
	age<40	Lab	1788	F(1,149)=2.179,	771	F(1,149)=0.902,	0.959	F(1,67)=0.857,	2.56	z=0.307	4	z=0.639		
		MTurk	1901	p=0.142	966	p=0.344	0.899	p=0.358	2.96	p=0.759	4.04	p=0.523		
bubble cursor									point		b1		b3	
	all age	Lab	*1331*	F(1,352)=0.728,	*257*	F(1,352)=2.878	*0.413*	F(1,233)=1.927,	7.58	z=0.201	2.33	z=0.036	1.08	z=0.251
	all mspf	MTurk	*1371*	p=0.394	*318*	p=0.091	*0.347*	p=0.166	8.34	p=0.841	3.4	p=0.971	1.77	p=0.802
	age<35	Lab	1171	F(1,77)=1.957,	195	F(1,77)=3.598,	0.418	F(1,42)=3.347,	8.44	z=0.222	1.78	z=0.78	1.22	z=0.805
	mspf<8	MTurk	1228	p=0.166	271	p=0.062	0.319	p=0.074	9.99	p=0.824	3.81	p=0.436	2.11	p=0.421

Table 2: Task completion times, individual standard deviations, utilization, and number of errors in all experiments. Except for errors, all reported values are least-squares means adjusted for device and gender. Values in *italics* are additionally adjusted for age and computer performance. None of the differences are statistically significant.

All experiments combined

We repeated the ANOVA analyses using the combined data from all experiments. The advantage of combining the data from all experiments is the increase in power of the test, making it possible to show statistically significant results for even smaller differences. However, despite the increased power, we did not observe a significant effect of setting on speed ($F(1,353) = 1.607, p = 0.206$), consistency ($F(1,353) = 1.509, p = 0.22$), or utilization ($F(1,353) = 2.037, p = 0.154$).

Outliers

In total, after the removal of participants with self-reported medical conditions or technical difficulties, 7 out of 290 (2.4%) Turkers and 1 out of 36 (2.8%) lab participants were classified as extreme outliers and excluded from the analysis across the three experiments.

Reproducibility

We run all the previously reported MTurk experiments on weekdays in the evenings at 7pm EDT (4pm PDT). The make up of the Turker population is known to change over time. To investigate if it would have impact on the reproducibility of the results, we rerun the Split Interface experiment on a weekday morning at 11am EDT (8am PDT). There were, indeed, substantial differences between the populations that participated in the two instances of the experiment: the morning population was 74% female, with median age of 43, and with 63% of the participants using a mouse and 27% using a touchpad. In contrast, the evening population was only 57% female, with median age of 32, and with 86% of the participants using a mouse and 14% using a touchpad. These differences had substantial impact on the average performance, but these performance differences became negligible and non-significant once age and device were introduced as factors into the analyses.

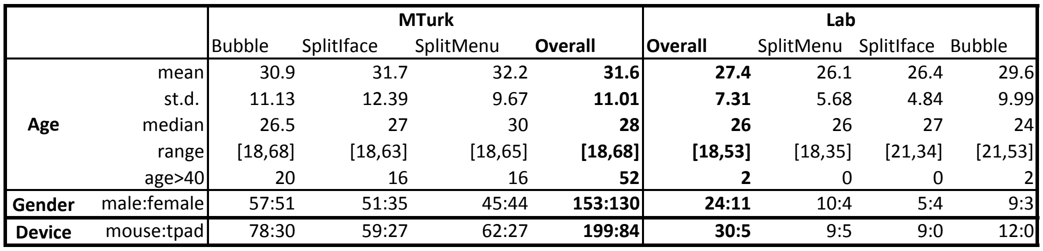
Demographics

Across the three experiments, we analyzed data from 283 MTurk participants (US-based, 95% approval rate, minimum of 500 HITs) aged 18–68, M=28. For the lab studies we analyzed data from 35 participants aged 18–53, M=26. Table 3 summarizes the demographics of the two populations. In lab, the age group 45–65 was underrepresented.

DISCUSSION

In all three experiments we have conducted, the data collected in lab and the data collected remotely with the participants recruited via MTurk both resulted in the same conclusions of the statistical comparisons and showed very similar effect sizes. We also did not observe any significant differences in the absolute measurements of task completion times, error rates, consistency of performance, or the rates of utilization of novel interface mechanisms between the lab and MTurk populations. Further, the results of the Split Menus and Split Interface experiments show that the Turkers did as well as lab-based participants on tasks where efficient operation of the user interface required some amount of cognitive engagement. We also found MTurk to be reliable in that when we repeated the same experiment at two different times of the day, we got nearly identical results after correcting for different age and input device distributions. We also detected very few outliers in our data: only 7 out of 290 participants (or 2.4%) had to be removed from the analysis across the three experiments. This is in contrast to other researchers who ended up discarding up to half of their MTurk-based data [8].

Although we designed our experiments as within-subjects comparisons, the close match in the absolute values of the measurements between the lab-based and online populations, as well as the lack of significant differences when the same measurements were repeated multiple times online, suggest that between-subjects designs could also be conducted robustly on MTurk.



		MTurk				Lab			
		Bubble	SplitIface	SplitMenu	**Overall**	**Overall**	SplitMenu	SplitIface	Bubble
Age	mean	30.9	31.7	32.2	**31.6**	**27.4**	26.1	26.4	29.6
	st.d.	11.13	12.39	9.67	**11.01**	**7.31**	5.68	4.84	9.99
	median	26.5	27	30	**28**	**26**	26	27	24
	range	[18,68]	[18,63]	[18,65]	**[18,68]**	**[18,53]**	[18,35]	[21,34]	[21,53]
	age>40	20	16	16	**52**	**2**	0	0	2
Gender	male:female	57:51	51:35	45:44	**153:130**	**24:11**	10:4	5:4	9:3
Device	mouse:tpad	78:30	59:27	62:27	**199:84**	**30:5**	9:5	9:0	12:0

Table 3: Demographics and pointing device of the final sample of participants, after excluding those classified as extreme outliers and those who reported a medical condition or a faulty device.

These positive results can be explained by the fact that participants were naturally motivated to complete tasks as quickly as possible, which aligned with the goals of the experiments. However, we cannot discount the possible effects of the novelty factor: a number of the MTurk participants commented spontaneously that our experiments were more enjoyable than typical tasks available on MTurk.

Our results suggest that remote performance evaluations of user interfaces can usefully be conducted with participants recruited via MTurk. However, because it is a novel methodology that hasn't been broadly validated, a small-scale lab-based validation may be a prudent choice for any new experiment.

Practical Considerations

Throughout this research, a number of practical considerations emerged that proved to be important for successful remote experimentation with Turkers. Although a number of these considerations have already been mentioned in earlier sections, we synthesize them all below.

- **Ensuring instruction comprehension.** During our preliminary studies, we saw ample evidence that a fraction of the Turkers did not understand the novel capabilities afforded by the user interface mechanisms we were testing. For example, some participants always brought the center of the Bubble Cursor over the target, while others never took advantage of the adaptive toolbar in Split Interfaces. To ensure that all participants understood what each interface afforded, we required that the participants used the novel capability of the UI being tested at least once during the practice block. If they didn't, they were shown the instructions again and were forced to repeat the practice block. We made it clear to participants that they were free to use their own judgement whether to take advantage of the novel UI capabilities during the actual experimental blocks. The comparable utilization rates between lab- and MTurk-based populations suggest that these interventions were effective.
- **Accounting for age- and input device–related differences.** Both age and the input device impact performance on input tasks such as pointing and text entry. While controlling for the diversity of age and input devices was not an issue for demonstrating within-subjects effects, between-subjects comparisons may not be reliable unless these factors are accounted for either when allocating participants to conditions or during the analysis. Because the make up of the MTurk work force changes on an hour-by-hour basis, running all conditions simultaneously is a common design choice that further contributes to the reliability of the results.
- **Robust outlier detection.** Some of the outliers we observed were more than an order of magnitude different from the typical performance. Such extreme outliers substantially impact the estimates of mean and standard deviation. Thus, the popular method of excluding values more than 2 standard deviations away from the mean may not be reliable. Instead, we have used a method based on the inter-quartile ranges, which is much more robust to very extreme outliers [4].
- **Implementation issues.** Some experiments, like the Bubble Cursor evaluation presented in this paper, will be sensitive to the hardware and software performance on the participants' computers. We both included automatic checks for detecting slow performing hardware/software configurations and we logged the actual time taken to render critical interface updates. Further, different participants will have different network connectivity. Standard techniques, such as preloading media, helped ensure that network performance was not a factor during the actual experiments.
- **Encouraging honest reporting of problems.** Twenty-one Turkers who completed our experiments reported having either a disability or a technical difficulty that might have impacted their performance during the experiment. We encouraged honest reporting of such problems by assuring participants that it would not affect their eligibility to receive compensation for participating in the study. While we didn't analyze the data from these participants, it is conceivable that not including them in the analysis contributed positively to the overall high quality of the data and the low outlier rates.

CONCLUSION

We have performed three distinct experiments both in a controlled laboratory setting and remotely with unsupervised online participants recruited via Amazon's Mechanical Turk (MTurk). The analysis of our results did not yield any evidence of significant or substantial differences in the data collected in the two settings: All statistically significant results detected in lab were also observed on MTurk, the effect sizes were similar, and there were no significant differences in the raw task completion times, error rates, measures of performance consistency, or the rates of utilization of the novel interface mechanisms introduced in each experiment. Repeated measurements performed on MTurk at different times of the day showed changes in the demographics of the MTurk population, but yielded similar results after correcting for different age and input device distributions observed at different times.

These results provide evidence that MTurk can be a useful platform for conducting experiments that require participants to perform largely mechanical user interface tasks and where the primary measures of interest are task completion times and error rates.

We have also highlighted a number of practical challenges — and our solutions to them — related to instruction comprehension, accounting for age- and device-related differences in performance, techniques for robust outlier detection, as well as the effects of software, hardware, and network performance on participants' experience.

Compared to lab-based methods, conducting online experiments with participants recruited through micro task labor markets such as Amazon's Mechanical Turk can enable larger-scale experimentation, access to a more diverse subject pool, and a faster experimental revision cycle. Although this is a novel methodology that has not been broadly validated, we believe it can provide a valuable complement to the existing approaches.

ONLINE APPENDIX

To enable others to validate and extend our results, the data set and more information about the design of our experiments can be found at http://iis.seas.harvard.edu/resources/.

ACKNOWLEDGEMENTS

Steven Komarov was supported by the National Science Foundation under fellowship number DGE-1144152. Katharina Reinecke was supported by the Swiss National Science Foundation under fellowship number PBZHP2-135922. We thank Mike Terry and the CHI reviewers for helpful feedback on earlier versions of this manuscript.

REFERENCES

1. Callison-Burch, C. 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*, EMNLP '09, Association for Computational Linguistics (Stroudsburg, PA, USA, 2009), 286–295.
2. Chandler, D., and Kapelner, A. Breaking monotony with meaning: Motivation in crowdsourcing markets. Working Paper, May 2010.
3. Cole, F., Sanik, K., DeCarlo, D., Finkelstein, A., Funkhouser, T., Rusinkiewicz, S., and Singh, M. How well do line drawings depict shape? *SIGGRAPH '09: SIGGRAPH 2009 papers* (July 2009).
4. Devore, J. *Probability and Statistics for Engineering and the Sciences*, seventh ed. Thomson Higher Education, 2008.
5. Gajos, K. Z., Czerwinski, M., Tan, D. S., and Weld, D. S. Exploring the design space for adaptive graphical user interfaces. In *AVI '06: Proceedings of the working conference on Advanced visual interfaces*, ACM Press (New York, NY, USA, 2006), 201–208.
6. Gajos, K. Z., Everitt, K., Tan, D. S., Czerwinski, M., and Weld, D. S. Predictability and accuracy in adaptive user interfaces. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ACM (New York, NY, USA, 2008), 1271–1274.
7. Grossman, T., and Balakrishnan, R. The bubble cursor: enhancing target acquisition by dynamic resizing of the cursor's activation area. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '05, ACM (New York, NY, USA, 2005), 281–290.
8. Heer, J., and Bostock, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, ACM (New York, NY, USA, 2010), 203–212.
9. Horton, J. J., Rand, D. G., and Zeckhauser, R. J. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* (2011).
10. Kapelner, A., and Chandler, D. Preventing Satisficing in Online Surveys: A "Kapcha" to Ensure Higher Quality Data. In *CrowdConf* (2010).
11. Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with mechanical turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, ACM (New York, NY, USA, 2008), 453–456.
12. Kong, N., Heer, J., and Agrawala, M. Perceptual Guidelines for Creating Rectangular Treemaps. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 990–998.
13. Little, G., Chilton, L., Goldman, M., and Miller, R. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, ACM (2010), 57–66.
14. Mao, A., Chen, Y., Gajos, K., Parkes, D., Procaccia, A., and Zhang, H. Turkserver: Enabling synchronous and longitudinal online experiments. In *Proceedings of HCOMP'12* (2012).
15. Mason, W., and Suri, S. Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods* (2010), 1–23.
16. Mason, W., and Watts, D. Financial incentives and the "performance of crowds". *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation* (June 2009).
17. Noronha, J., Hysen, E., Zhang, H., and Gajos, K. Z. Platemate: Crowdsourcing nutrition analysis from food photographs. In *Proceedings of UIST'11* (2011).
18. Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., and Biewald, L. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Proceedings of HCOMP'11* (2011).
19. Oppenheimer, D., Meyvis, T., and Davidenko, N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology 45*, 4 (2009), 867–872.
20. Paolacci, G., Chandler, J., and Ipeirotis, P. Running experiments on amazon mechanical turk. *Judgment and Decision Making 5*, 5 (2010), 411–419.
21. Prelec, D. A Bayesian Truth Serum for Subjective Data. *Science 306*, 5695 (Oct. 2004), 462–466.
22. Rand, D. G. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of theoretical biology* (2012).
23. Rzeszotarski, J. M., and Kittur, A. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, ACM (New York, NY, USA, 2011).
24. Schmidt, L. Crowdsourcing for human subjects research. In *CrowdConf* (2010).
25. Sears, A., and Shneiderman, B. Split menus: effectively using selection frequency to organize menus. *ACM Trans. Comput.-Hum. Interact. 1*, 1 (1994), 27–51.
26. Shaw, A. D., Horton, J. J., and Chen, D. L. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, ACM (New York, NY, USA, 2011), 275–284.
27. Suri, S., and Watts, D. Cooperation and contagion in networked public goods experiments. *Arxiv preprint arXiv10081276* (2010).