Differences in Crowdsourced vs. Lab-based Mobile and Desktop Input Performance Data

Leah Findlater[1], Joan Zhang[2], Jon E. Froehlich[2], Karyn Moffatt[3]
[1,2]Human-Computer Interaction Lab
[1]College of Information Studies | [2]Dept. of Computer Science
University of Maryland, College Park, MD
leahkf@umd.edu, joan.r.zhang@gmail.com, jonf@cs.umd.edu

[3]ACT Research Group
School of Information Studies
McGill University, Montreal, QC
karyn.moffatt@mcgill.ca

ABSTRACT
Research on the viability of using crowdsourcing for HCI performance experiments has concluded that online results are similar to those achieved in the lab—at least for desktop interactions. However, mobile devices, the most popular form of online access today, may be more problematic due to variability in the user's posture and in movement of the device. To assess this possibility, we conducted two experiments with 30 lab-based and 303 crowdsourced participants using basic mouse and touchscreen tasks. Our findings show that: (1) separately analyzing the crowd and lab data yields different study conclusions—touchscreen input was significantly less error prone than mouse input in the lab but *more* error prone online; (2) age-matched crowdsourced participants were significantly faster and less accurate than their lab-based counterparts, contrasting past work; (3) variability in mobile device movement and orientation increased as experimenter control decreased—a potential factor affecting the touchscreen error differences. This study cautions against assuming that crowdsourced data for performance experiments will directly reflect lab-based data, particularly for mobile devices.

Author Keywords
Human performance; crowdsourcing; input devices; mobile.

ACM Classification Keywords
H.5.2. Input devices and strategies (*e.g.*, mouse, touchscreen)

INTRODUCTION
Online crowdsourcing platforms such as Amazon's Mechanical Turk (MTurk) are increasingly popular for conducting fast, relatively inexpensive user studies with a large number of participants. A substantial body of work has examined and discussed the viability of crowdsourcing human-subjects studies—across fields as diverse as human-computer interaction (HCI) [19], psychology [4], behavioral economics [15], and political science [2].

CHI 2017, May 06-11, 2017, Denver, CO, USA
 ISBN 978-1-4503-4655-9/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025820

Critical to understanding this viability are experiments that compare crowdsourced data to lab-based data. A common HCI focus that has received relatively little attention from this perspective is performance evaluation of input devices and interaction techniques. Such studies are typically conducted in controlled lab settings and performance is measured in milliseconds, making even short bursts of inattention problematic. Komarov *et al.* [21] recently compared data collected in the lab to data collected through MTurk for three desktop-based interaction techniques: adaptive split menus [29], split interfaces [11], and bubble cursor [12]. The study found no evidence of "significant or substantial differences" between the lab and crowdsourced data, suggesting that crowdsourcing may indeed be a valid means of conducting HCI performance studies.

However, at least two important questions remain. First, it is unclear whether Komarov *et al.*'s [21] findings extend to other types of devices. In particular, mobile devices may make remote data collection more problematic due to movement of the device itself and variability in the user's posture. Given that mobile device internet usage has now eclipsed that of desktops [8], this question is especially pertinent to online data collection. Second, the lab-based sample sizes employed by Komarov *et al.* were relatively small (N=10 to 14), selected to match sample sizes in the original experiments that were being replicated, but likely resulting in low statistical power when directly comparing the lab-based and crowdsourced groups. The question of power is especially crucial here as the implications of their study hinge on the *lack* of significant differences found.

To address these questions, we conducted two experiments comparing crowdsourced and lab-based data. For the first experiment, we recruited 202 MTurk participants and 30 lab-based participants. Participants completed a set of basic input tasks (crossing, dragging, pointing, and steering) using an indirect pointing device (*e.g.*, a mouse) and/or a touchscreen tablet. This experiment thus builds on Komarov *et al.*'s [21] study by including a mobile device and a different set of interaction techniques, but also provides a larger sample size for revisiting findings from that study. Our findings show that, in contrast to Komarov *et al.* [21], (1) separately analyzing the crowdsourced and in-person data yielded different study conclusions—touchscreen input was significantly less error prone than mouse input in the lab but *more* error prone online, and (2)

by directly comparing the two data sources, age-matched crowdsourced participants were significantly faster and less accurate than their lab-based counterparts.

The second experiment delved more deeply into a potential factor affecting the touchscreen accuracy results: whether degrees of experimental control affect the physical orientation and movement of the mobile device itself. In the first experiment, we had asked participants to lay their tablet flat during tasks. In this second experiment, we recruited 101 new crowdsourced touchscreen participants who were *not* given that instruction and compared them to the original participants for whom we had touchscreen data. Extraneous device movement and rotation of the device away from a horizontal position both increased significantly as experimental control decreased—from the lab setting to the remote setting to removing the placement instruction.

The primary contributions of this paper are: (1) empirical evidence showing that there are significant differences in performance between crowdsourced and lab-based study participants, in contrast to previous work [21], and (2) a characterization of how variation in experimental control (in terms of environment and instruction) affects movement and orientation of a mobile device—an important factor potentially impacting remote collection of mobile input data. These contributions caution against assuming that performance data collected from paid crowdworkers will directly reflect data from a more controlled setting, both for mouse-based input and, particularly, for mobile input. We discuss the potential implications of these findings.

RELATED WORK

Crowdsourcing versus Traditional Lab Studies

Studies on the viability of using crowdsourcing as an alternative to traditional lab-based experiments have spanned fields from HCI to political science. Burhmester *et al.* [4], for example, analyzed the demographic makeup of a large MTurk sample and compared test-retest reliability of psychometric data, concluding that the MTurk data met acceptable standards. Horton *et al.* [15] also found they could replicate a set of behavioral economics experiments using MTurk and argued that these types of online experiments have high internal validity. Similar conclusions were drawn by Berinsky *et al.* by replicating a series of political science experiments [2]. In HCI, an important comparison of lab versus crowdsourced data collection is Heer and Bostock's [13] replication of visual perception studies using MTurk. Although the crowdsourced data was more variable, it closely matched previous lab-based results and design implications were identical from the two data sources. Their study also showed that using qualification tests and verifiable questions results in high quality data.

The closest work to ours, and upon which we build, comes from Komarov *et al.* [21], who compared crowdsourced and lab-based performance data for three desktop-based interaction techniques. For each of the three experiments, they (1) *separately* analyzed the crowdsourced and lab-based data to determine if there were any differences in the study conclusions reached from the two sources (*e.g.*, is Condition A faster than Condition B for both sources), and (2) *directly compared* the sources to identify any statistically significant differences between them (*e.g.*, are crowdsourced participants consistently faster than lab-based participants). As already mentioned, based on these two types of analyses, the study's conclusion was that there were no significant or substantial differences between the two sources of data. We follow Komarov *et al.*'s method, extending it to a new device and different tasks, and highlight where our findings contrast theirs.

Finally, while there is relatively little work on understanding the viability of employing crowdworkers for performance experiments, a number of studies have recently increased the scale at which HCI performance data is collected, by creatively designing and deploying online apps through the Apple App Store or Google Play Store. Examples include collecting touch input data [14] or text entry performance data [28] through mobile games. However, this type of data collection differs in terms of participant incentives and goals (*i.e.*, to have fun or be entertained) from our focus, which is *paid* crowdwork [20].

Note also that work exists on crowdsourcing design critiques (e.g., [34])—subjective feedback on user interface designs—which contrasts our focus on input experiments that require precise performance measurements.

Who are Mechanical Turk Workers?

Diversity is seen as a strength of the MTurk participant pool [24,25], although this diversity can be affected by factors such as pay, task complexity, and sampling time [25]. Paolacci and Chandler [25] surveyed studies characterizing MTurk samples and concluded that workers were dominated by US and Indian residents, tended to be younger, overeducated, underemployed, less religious, and more liberal than the population as a whole. Other work has reported that MTurk samples tend to have more females than males [24], at least for US samples [17,18]. In an update of their 2010 paper [27] on MTurk demographics, Silberman *et al.* [30] recently noted that demographics have shifted in the past five years and that "professional Turkers" now complete most tasks in the system and have a stronger incentive than other workers to seek out high paying tasks and perform them well. As such, it may be timely to revisit findings from earlier crowdsourcing studies.

Quality Concerns with Crowdsourced Data

Obtaining and assessing high-quality data is an important challenge for crowdsourcing [1,20]. Quality issues range from cheating behaviors such as arbitrarily selecting answers or copying answers from elsewhere [9] to less malicious behavior such as not paying close attention [21]. In early work experimenting with crowdsourcing for subjective data collection in HCI, Kittur *et al.* [19] recommended that quality could be improved by including explicitly verifiable questions in the task and making cheating equally effortful to completing the task accurately.

Since then, a number of quality control mechanisms have become popular, such as redundancy, where multiple workers complete the same task, reputation systems, ground truth seeding, statistical filtering, and expert review [26]. Providing feedback through "shepherding" can also lead to higher quality work [7]. At a high level, these approaches can be grouped into up-front task design approaches versus posthoc result analysis approaches [20].

Many popular quality control approaches, however, are not relevant to performance studies such as ours; for example, ground truth seeding (using gold standard data) and redundancy do not have a direct analog. Instead, careful task design, comprehensible instructions, and statistical outlier detection (as used in [21]) are particularly important—we employ these in our study.

EXPERIMENT 1: COMPARING CROWDSOURCED AND LAB-BASED INPUT PERFORMANCE

To compare crowdsourced mobile input performance data to that collected in the lab, and to determine whether this comparison is different from desktop data, we recruited 30 lab participants and 202 crowd participants. We build on Komarov *et al.*'s [21] study by largely following their method, but extend it to mobile devices and a different set of interaction techniques (in our case, crossing, dragging, pointing, steering). Our larger sample size is also useful for revisiting their study conclusions comparing crowdsourced and lab-based data with desktop interaction techniques.

Method

Apparatus and Tasks

The experiment testbed was built in JavaScript, PHP, and HTML and tested in major browsers (Chrome, Safari, Internet Explorer, Firefox) on desktop and laptop computers and in Safari on Apple iOS. In the lab, the testbed was loaded in Safari on an Apple iPad 3 (9.7" screen) and in Chrome on an Apple laptop running Mac OS X El Capitan connected to a 1280×1024-resolution external monitor and a Logitech M310 wireless optical mouse. Crowdsourced participants could use their preferred browser on a laptop or desktop, but had to use an iPad for touchscreen input due to inconsistent performance of some Android tablets. The software logged touch and mouse input, as well as accelerometer and gyroscope data from the iPad.

The testbed guided participants through four basic input tasks: crossing, dragging, pointing, and steering (Figure 1). The tasks were implemented based on the ISO 9241-9 circle 2-D Fitts' law task standard [16]. To attain a range of indexes of difficulty, we fully crossed amplitudes (A) of {250px, 500px} with widths (W) of {32px, 64px, 96px}. We then removed the combination (A=500px, W=96px) because it did not fit on the iPad screen with enough padding to allow participants to overshoot the target by a distance of $2W$ (the iPad is only 768px wide and the task canvas had to be square for the ISO task). Index of difficulty (ID) is the ratio between the distance to the target and the target's width, with higher ID values indicating

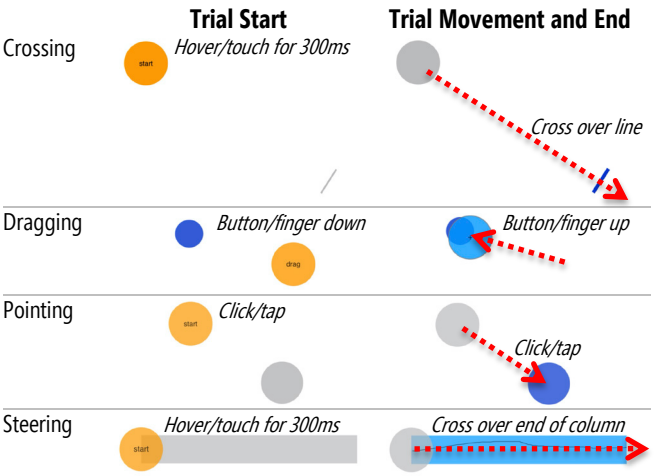

Figure 1. Example trials (cropped screenshots) from the four tasks participants completed with the mouse and/or touchscreen: crossing, dragging, pointing, and steering. As seen here, trials varied in terms of angle of movement, based on ISO 9241-9 [16]. To start a trial, the participant had to activate the start area target (orange circle) by clicking or tapping in the dragging and pointing tasks or by hovering or holding for 300ms in the crossing and steering tasks.

greater input difficulty [31]: $ID = \log_2(A/W+1)$. The five $A \times W$ combinations thus provide an ID range of 1.9–4.1. On a 9.7" iPad, the smallest width, 32px, corresponds to a 6mm-wide target.

Participants

Table 1 shows demographics and group sizes for the three participant groups recruited for this experiment: crowdsourced mouse, crowdsourced touchscreen, and lab-based. For the crowdsourced mouse group (N=101), 85 participants reported using an optical mouse to complete the study, 13 used a touchpad, and three used a mechanical mouse. Crowdsourced touchscreen participants were required to have access to an Apple iPad to participate. Crowdsourced participants were recruited through MTurk and were paid $4 to complete the study, while lab-based participants were recruited through campus mailing lists and compensated $15. The difference in compensation reflects both that the procedure for lab-based participants was twice as long as for crowdsourced participants (the former completed both touchscreen and mouse conditions) and that lab-based participants had to travel to our lab.

N	Age	Gender	Experience
Crowdsourced mouse			
101	*M*=35.4 *range*=19–69 *SD*=11.0	41 women 59 men 1 other	99 daily computer users
Crowdsourced touchscreen			
101	*M*=32.0 *range*=19–59 *SD*=8.5	51 women 50 men	90 daily touchscreen users
Lab-based			
30	*M*=19.9 *range*=18–29 *SD* = 2.1	17 women 13 men	30 daily computer users 29 daily touchscreen users

Table 1. Participant groups for Experiment 1.

Procedure

The crowdsourced and lab-based participants completed an almost identical study procedure. The main difference, in addition to the presence of an experimenter and use of a quiet room for lab-based participants, was that crowdsourced participants only used one of the input devices (mouse or touchscreen), whereas lab-based participants used both devices (*i.e.*, a fully within-subjects design). The choice to use a within-subjects design in the lab reflects standard practice in HCI performance experiments, while the between-subjects design for crowdsourced participants reflects the piecemeal participation that is common with crowdsourced data collection. This difference between the crowdsourced and lab-based data collection is considered later in our analyses.

Crowdsourced procedure. The study was advertised as taking 20-30 minutes, though some participants completed it more quickly. After viewing and agreeing to consent information, participants completed four basic input tasks (crossing, pointing, dragging, steering), presented in random order.[1] Touchscreen participants were asked to place the iPad on a flat surface such as a table. Each task included five practice trials, which participants could repeat once if they chose, followed by 55 test trials presented in random order (11 trials for each $A \times W$ combination). Each trial began by activating a circular start target by tapping/clicking (for the pointing and dragging tasks) or by holding/hovering for 300ms (for the crossing and steering tasks). Spatial outlier trials were automatically redone by appending them to the end of the set of trials. These outliers were defined based on [23] as trials where (1) the actual movement distance was less than half of A or (2) the endpoint of the trial (*e.g.*, mouse click) was more than $2W$ away from the ideal endpoint (*i.e.*, the target center).

The testbed software enforced the use of an iPad as opposed to other touchscreen devices and required that the iPad be used in portrait orientation. A modal dialog was displayed on the screen with a warning message if the orientation changed or if more than three fingers were detected at once. A warning message was also displayed for both the mouse and touchscreen conditions if no user activity was detected for 30 seconds. After dismissing one of these warning dialogs, the current trial restarted. A break was offered halfway through each task. Between tasks, short questionnaires collected demographic data and device use.

A strength of crowdsourcing is that the computer setups will be more ecologically valid than in the lab, varying, for example, in display size and mouse model. Specifically for iPads, three sizes were available: 7.9", 9.7" (the standard size), and 12.9". Physical target sizes and distances were the same for the 9.7" and 12.9" versions, while the smallest size proportionately scaled down output such that a 6mm target on the medium iPad was 4.9mm on the small iPad. Apple does not make it possible to programmatically distinguish between the small and medium iPads using JavaScript/HTML[2] and we did not originally ask participants to self-report device size. Based on Fitts' Law, there should theoretically be no impact on input performance for our tasks: due to the proportional scaling, the indexes of difficulty are identical for both sizes; however, some studies suggest that display size *can* impact performance [22]. To assess whether iPad size impacted performance, we collected another set of crowdsourced touchscreen data from 37 small iPad users and 61 medium iPad users. There were no significant differences between the two devices in terms of trial completion time (unpaired t-test: $t_{=96} = 0.46$, $p = .647$, $d = 0.08$) and error rate (Mann Whitney U test: $Z = 0.58$, $p = .562$, $r = 0.06$) across the four study tasks; effect sizes were also close to zero.

Lab-based procedure. Study sessions lasted up to one hour, longer than for crowdsourced participants due to the within-subjects design used in the lab: participants completed all four tasks with both the mouse and the touchscreen tablet. Input device (mouse, touchscreen) was fully counterbalanced, while, again, tasks were randomly ordered for each device. For the touchscreen, the iPad was placed flat on a table. For each input device, the experimenter loaded the study webpage in the browser before asking the participant to independently complete the same procedure as described for crowdsourced participants.

Dataset and Analysis

The dataset includes 44,440 trials from crowdsourced participants and 13,200 trials from lab-based participants. To manage quality issues that can arise with crowdsourced data, we identified: (1) trial-level outliers that could have resulted from momentary distraction (*e.g.*, 45 seconds to complete a trial that had previously been completed in 5-10 seconds) and (2) participant-level outliers that could have resulted from a more systematic confound such as watching television while completing the study. We used the interquartile range (IQR) method, which is more robust than mean-based outlier approaches to extreme outliers, like the trial-level example above. An extreme outlier trial is defined as being more than 3×IQR above the third quartile or below the second quartile [6]. For trials, this calculation was made within each $A \times W$ condition for each task for each participant, removing 1.0% of trials across all participant groups. Participant-level outliers were computed based on average trial completion time across all tasks in the respective participant group. The exact numbers of participant-level outliers are reported in Results.

Our main measures were trial completion time and error rate. We checked the normality assumption required of parametric tests using Q-Q plots and Shapiro-Wilk tests. Both time and error rate violated this assumption. For time, we log-transformed the data to meet the normality

[1] With the touchscreen only, participants completed two additional tasks (pinch/zoom), but these appeared only *after* the four primary tasks and so do not affect performance on the earlier tasks. Because they were not part of the primary experimental design, we do not report on them here.

[2] http://stackoverflow.com/questions/13248493/detect-ipad-mini-in-html5

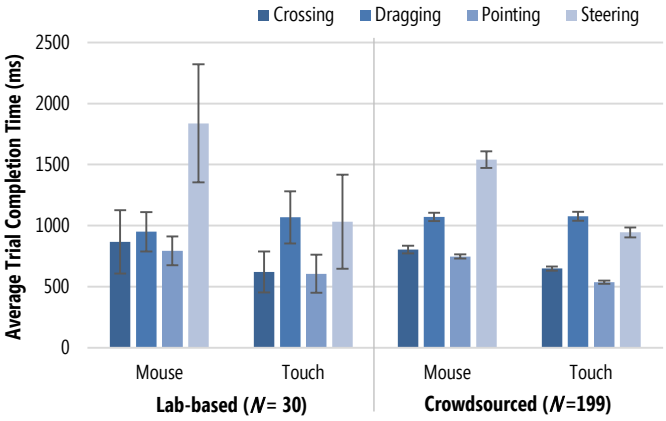

Figure 2. Average trial completion time for all participants in Experiment 1. Lower values are better and error bars show standard error. Analyzing the crowdsourced and lab-based data separately results in identical conclusions.

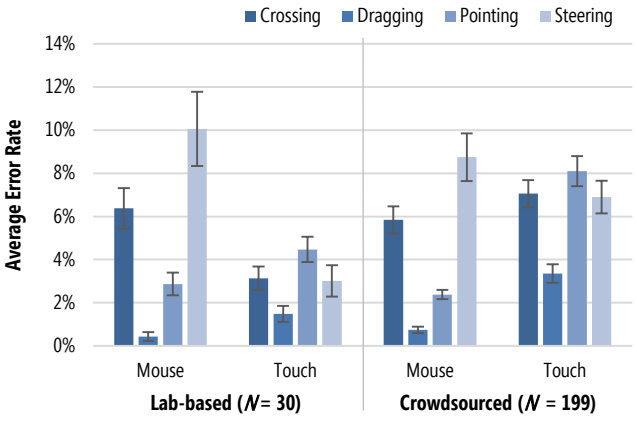

Figure 3. Average error rate results for all participants in Experiment 1. Lower values are better and error bars show standard error. Error rates were lower with the touchscreen compared to the mouse for lab-based participants, while the *opposite* was true of crowdsourced participants.

assumption before applying ANOVAs and other parametric tests. For error rate, we employed non-parametric ANOVAs with Aligned Rank Transform (ART) [32], with Mann Whitney U or Wilcoxon signed ranks tests for posthoc comparisons. For all ANOVAs, when the degrees of freedom are fractional, a Greenhouse-Geisser adjustment has been applied. Bonferroni adjustments were applied to all posthoc pairwise comparisons. Finally, where non-significant effects are reported, observed power is included to help with interpretation.

Results

Following Komarov *et al.*'s [21] approach, we report on two analyses: (1) Examining the crowdsourced and lab-based data *separately* to assess possible differences in study conclusions reached from each data source—for example, is the touchscreen faster than the mouse for one dataset but not the other? (2) *Directly comparing* the two sources of data to determine whether statistically significant differences exist between them—for example, are participants in one dataset consistently faster than the other across all experimental conditions? For simplicity, we refer to all indirect pointing devices as "mouse", though some crowdsourced participants used other devices (see above).

Separately Analyzing Study Conclusions

After removing participant-level outliers in each of the three participant groups, this analysis includes 99 crowdsourced mouse participants, 100 crowdsourced touchscreen participants, and 30 lab-based participants. First, we analyze the lab-based data to create a baseline for comparison. Then, we analyze the crowdsourced data and compare its conclusions with those from the lab-based data.

Lab-based data. Lab-based participants were faster and more accurate with the touchscreen than the mouse, which is in line with previous findings in the lab (*e.g.*, [10]). These overall patterns are shown in Figures 2 and 3. We computed separate 2×4 repeated measures ANOVAs for trial completion time and error rate, with factors of *Device* (2 levels: mouse or touchscreen) and *Task* (4 levels: crossing, dragging, pointing, steering); in the case of error rate, we used an ANOVA with ART.

Trial Completion Time. Average trial time was 831ms (SD = 329) with the touchscreen compared to 1,112ms (SD = 513) with the mouse. This difference was statistically significant, as shown by a main effect of *Device* ($F_{1,29}$ = 88.06, p < .001, η_p^2 = .75). *Task* also significantly impacted trial time (main effect: $F_{2.13,61.86}$ = 129.93, p < .001, η_p^2 = .82) and some tasks were impacted differently depending on what device was used (*Task* × *Device* interaction effect: $F_{3,87}$ = 66.86, p < .001, η_p^2 = .49).

Based on the interaction effect, we conducted posthoc pairwise comparisons between devices for each task. These showed that the touchscreen was faster than the mouse for crossing, pointing, and steering, but the mouse was faster for dragging (all p < .05). For example, the average time for steering trials decreased from 1,837ms (SD = 483) with the mouse to 1,032ms (SD = 386) with the touchscreen. Dragging, in contrast, went from 950ms (SD = 161) with the mouse to 1,068ms (SD = 214) with the touchscreen.

Error Rate. Error rates (Figure 3) were also lower with the touchscreen than with the mouse, at on average 3.0% (SD = 3.3) versus 4.9% (SD = 6.6) respectively. This difference was statistically significant (main effect of *Device*: $F_{1,29}$ = 27.90, p < .001, η_p^2 = .49). As with trial time, error rates were also significantly different based on task (main effect of *Task*: $F_{3,87}$ = 28.67, p < .001, η_p^2 = .50), and some tasks tended to be more error prone with one device than the other (*Task* × *Device* interaction effect: $F_{2.43,70.52}$ = 37.34, p < .001, η_p^2 = .56).

Posthoc pairwise comparisons based on the interaction effect showed that participants made significantly fewer errors with the touchscreen for crossing and steering (both p < .05), but made significantly more errors with the touchscreen for pointing (p < .05); there was no difference between the two devices for dragging. These differences

were most notable with steering, where there were approximately three times as many errors with the mouse (M = 10.1%, SD = 9.4) as with the touchscreen (M = 3.0%, SD = 4.0), and crossing, where there were twice as many errors with the mouse (M = 6.4%, SD = 5.1) as with the touchscreen (M = 3.1%, SD = 3.0).

Finally, for the pointing task only, we computed Fitts' law models for all participants using bivariate endpoint deviation [33]. Average model fits (Pearson r) were higher with the mouse at r = .95 (SD = .05) than with the touchscreen at r = .90 (SD = .08). This difference was significant with a Mann-Whitney U test (U = 265, Z = 2.73, p = .006, r = 0.35), and is not surprising given that a standard Fitts' law model does not adjust for the "fat finger" problem on touchscreens [3]. Reflecting the speed and error rate results above, throughput was higher with the touchscreen (M = 5.1, SD = 1.2) than with the mouse (M = 3.6, SD = 0.4); this difference was statistically significant with a paired t-test (t_{29} = 7.18, p < .001, d = 1.72).

Summary. The lab-based conclusions were as expected based on past comparisons of basic input tasks between mouse and touchscreen devices (*e.g.*, [5,10]). The touchscreen was faster than the mouse for all tasks except dragging. The touchscreen also had lower error rates, most dramatically with steering where there were three times as many errors with the mouse than with the touchscreen.

Crowdsourced data. Conclusions from the crowdsourced data were similar to the lab-based data for trial completion time but not for error rate. We computed separate 2×4 repeated measures ANOVAs for time and error rate, with the between-subjects factor of *Device* (2 levels) and the within-subjects factor of *Task* (4 levels); again, in the case of error rate, this was an ANOVA with ART.

Trial Completion Time. The touchscreen was significantly faster than the mouse, at 801ms on average per trial (SD = 363) compared to 1,040ms (SD = 520) with the mouse (main effect of *Device*: $F_{1, 197}$ = 45.48, p < .001, η_p^2 = .19). The tasks also significantly impacted time (main effect of *Task*: $F_{2.59,510.37}$ = 359.83, p < .001, η_p^2 = .65). Again, this difference was more pronounced for some task and device combinations than others (*Task* × *Device* interaction effect: $F_{2.59,510.37}$ = 44.24, p < .001, η_p^2 = .18). Across tasks and devices, average trial time ranged from 537ms (SD = 128) for pointing on the touchscreen to 1,540ms (SD = 675) for steering with the mouse. Mirroring the conclusions for the lab-based data, pairwise comparisons between the two devices for each task showed that the touchscreen was faster than the mouse for all tasks but dragging, where the mouse was faster (all p < .05).

Error Rate. For error rates, however, the conclusions diverged from the lab-based data. Error rates were on average 6.4% (SD = 6.6) with the touchscreen compared to 4.5% (SD=7.1) with the mouse. Thus, while there was again a significant difference between the touchscreen and mouse input, the direction of that difference was the *opposite* to what it had been with the lab-based data: the touchscreen here was *less* accurate than the mouse, rather than more accurate (main effect of *Device*: $F_{1,197}$ = 13.19, p < .001, η_p^2 = .06). There were also significant effects of *Task* ($F_{2.88,566.30}$ = 66.69, p < .001, η_p^2 = .25) and *Task* × *Device* ($F_{2.79,549.62}$ = 13.10, p < .001, η_p^2 = .06).

Posthoc pairwise comparisons based on the interaction effect showed that dragging and pointing were significantly less accurate with the touchscreen than the mouse (both p < .05). Dragging error rates went from 0.7% (SD = 1.4) with the mouse to 3.3% (SD = 4.3) with the touchscreen, while pointing error rates more than tripled from 2.4% (SD = 2.2) with the mouse to 8.1% (SD = 7.0) with the touchscreen.

Finally, we again computed Fitts' law models for the pointing task only. On average, the model fits for mouse participants were r = .926 (SD = .079), whereas they were r = .898 (SD = .099) for touchscreen participants. Unlike with the lab-based data, this difference was not statistically significant with a Mann-Whitney U test (U = 4166, Z = 1.71, p = .087, r = 0.12). Throughput, however, was again significantly higher with the touchscreen (M = 5.5, SD = 1.2) than with the mouse (M = 4.1, SD = 0.7), as shown with an unpaired t-test (t_{197} = 10.28, p < .001, d = 1.46).

In summary, trial completion time conclusions here are consistent with the lab: the touchscreen is faster than the mouse for all tasks but dragging. However, the touchscreen here was *less* accurate than the mouse, which is the direct opposite of the lab-based conclusions.

Directly Comparing Crowdsourced vs. Lab-Based Data

Again following Komarov *et al.*'s [21] method, our second analysis directly compares the lab-based and crowdsourced data to determine whether there are any systematic differences between the two sources (*e.g.*, one group is consistently faster), in addition to the differences in study conclusions already seen. Rather than including all participants here, however, we roughly age-match the two groups. The lab-based group was younger than the crowdsourced group, which is unsurprising given that the lab-based sample, as is often the case, was drawn from a university campus. Indeed, the potential to obtain a more diverse participant sample is a strength of crowdsourcing studies [26,27]. However, because age affects performance with both mice and touchscreens (*e.g.*, [10]), age-matching participants provides a cleaner comparison of how the lab versus remote setting impacts performance. Further, and again to ensure a clean comparison, we only include data from the *first* device each lab participant used, to mirror the between-subjects design of the crowdsourced data.

Data and analysis. To accommodate the above goals, this analysis includes only 15 lab-based touchscreen participants and 15 lab-based mouse participants (*i.e.*, only the participants who had used each of those devices first during the study procedure). We then randomly selected 15

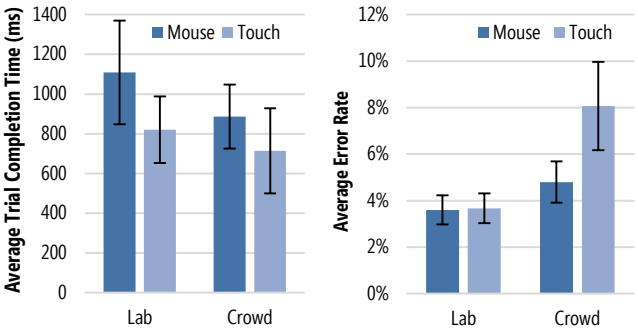

Figure 4. Average trial completion time (left) and error rate (right) for the direct comparison of lab-based participants to a roughly age-matched subset of crowdsourced participants in Experiment 1. (*N*=60 in total; error bars show standard error.)

touchscreen and 15 mouse participants who were younger than 30 years old from the crowdsourced group (the oldest lab-based participant was 29). There were no participant-level outliers in any of the participant groups included in this analysis. Finally, to simplify analysis and focus on differences between the two devices (mouse *vs.* touchscreen) and the two groups (crowdsourced *vs.* lab-based), we analyze trial completion time and error rate across all tasks rather than including *Task* as a factor. We also analyze Fitts' law models for the pointing task only.

Results. Overall, crowdsourced participants were faster and more error prone than lab-based participants; see Figure 4. We computed separate 2×2 ANOVAs (*Group* × *Device*) for each of these measures; ANOVA with ART in the case of error rate. Crowdsourced participants completed trials in 801ms on average (SD = 240), significantly faster than the 965ms (SD = 249) for lab-based participants (main effect of *Group*: $F_{1,56} = 8.83$, $p = .004$, $\eta_p^2 = .14$). Mirroring the device comparison results from the earlier analysis, the touchscreen was also faster overall than the mouse (main effect of *Device*: $F_{1,56} = 17.69$, $p < .001$, $\eta_p^2 = .24$). The *Group* × *Device* interaction effect was not significant ($F_{1,56} = .44$, $p > .05$, $\eta_p^2 = .01$, observed power = .10).

The average error rate for crowdsourced participants was 6.4% (SD = 5.9), which was almost double the 3.6% (SD = 2.4) error rate of lab-based participants (main effect of *Group*: $F_{1,56} = 5.72$, $p < .020$, $\eta_p^2 = .09$). Touchscreen error rates were particularly high for the crowdsourced participants, at 8.1% (SD = 7.4), compared to the lab-based participants, at 3.7% (SD = 2.5). However, neither the main effect of *Device* ($F_{1,56} = 1.22$, $p > .05$, $\eta_p^2 = .02$, observed power = .19) nor the *Group* × *Device* interaction effect ($F_{1,56} = .24$, $p > .05$, $\eta_p^2 < .01$, observed power = .08) were significant, perhaps due to the smaller sample size with this between-subjects experiment relative to our prior within-subjects analysis of the lab-based data.

Finally, we computed Fitts' law models for the pointing task. Crowdsourced participants had higher throughput (M = 5.1, SD = 1.3) than lab-based ones (M = 4.4, SD = 1.2), which a 2×2 ANOVA showed was significant (main effect of *Group*: $F_{1,56} = 7.70$, $p = .007$, $\eta_p^2 = .12$). This finding shows that the different speed-accuracy tradeoffs of the two groups were not comparable. The touchscreen also yielded higher throughput (M = 5.4, SD = 1.3) than the mouse (M = 4.0, SD = 0.6) (main effect of *Device*: $F_{1,56} = 32.80$, $p < .001$, $\eta_p^2 = .37$), while a *Group* × *Device* interaction effect was not significant ($F_{1,56} = 1.56$, $p > .05$, $\eta_p^2 < .03$, obs. power = .23). Model fits (r) were on average 0.94 (SD = 0.06) for the lab-based group and 0.94 (SD = 0.07) for the crowdsourced group. A 2×2 ANOVA with ART revealed no significant main or interaction effects on r (*Group*: $F_{1,56} = .14$, $p > .05$, $\eta_p^2 < .01$, obs. power = .07; *Device:* $F_{1,56} = 2.39$, $p > .05$, $\eta_p^2 = .04$, obs. power = .33; *Group* × *Device:* $F_{1,56} = .03$, $p > .05$, $\eta_p^2 < .01$, obs. power = .05).

Summary

Separately analyzing the crowdsourced and lab-based data showed that the error rate conclusions derived from the two datasets were in opposition. More specifically, touchscreen error rates were higher compared to the mouse with the crowdsourced data, while the opposite was true of the lab-based data. By directly comparing age-matched participants from both groups, we also found that crowdsourced participants made a different speed-accuracy tradeoff than lab-based participants: the former were faster and less accurate than the latter. Overall, our findings contrast Komarov *et al.*'s [21] study of desktop-based interaction techniques, which we return to in the Discussion section.

In the next experiment, we examine one potential reason for differences in error rates with the mobile device: variability of device orientation and extraneous device movement that may occur outside of the controlled lab setting.

EXPERIMENT 2: IN-DEPTH TOUCHSCREEN USE

A mobile touchscreen device can be placed or held at different angles and can move during use, which introduces a potential source of performance variability when moving from a controlled setting to crowdsourced data collection. To understand the extent of this variability and its relationship to input performance, we examine tablet orientation and device movement, using the touchscreen data from Experiment 1 along with a new set of 101 crowdsourced participants. In Experiment 1, crowdsourced participants were asked to place the tablet on a flat surface such as a table or desk (Figure 5), while all lab-based participants used the tablet on a desk. To assess more natural mobile touchscreen behavior from crowdsourced participants, this new set of participants completed the study *without* any device placement instruction.

Method

Experiment 2's method is largely the same as that used in Experiment 1. We highlight only the differences here.

Participants

In addition to the 101 crowdsourced *touchscreen* participants and 30 lab-based participants from Experiment 1, 101 new participants were recruited through MTurk and

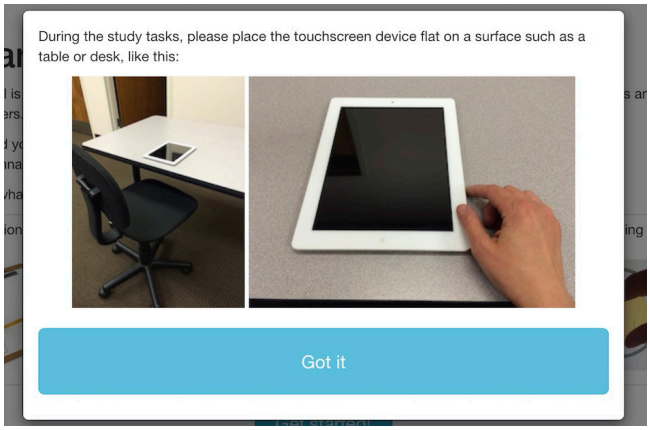

Figure 5. Device placement instructions for crowdsourced touchscreen participants in Experiment 1. This message was omitted for participants without instruction in Experiment 2.

were paid $4 to participate. This new set of participants was on average 33.8 years old (*range* 20–66, *SD* = 8.3) and consisted of 59 women and 42 men. All used an Apple iPad for the study and 93 reported at least daily touchscreen use.

Procedure

The procedure for the new participants was exactly the same as for crowdsourced touchscreen participants in Experiment 1, except that the new group was given no instruction on how to hold or place the device.

Data and Analysis

Of the new participants (*i.e.*, crowdsourced participants *without* instruction on how to hold the device), two used first generation iPads that did not have a gyroscope, and are thus excluded from the analysis. After removing an additional participant-level outlier, 98 participants remained; across participants, 1.2% of trials were removed as temporal outliers. Of the 101 crowdsourced touchscreen participants from Experiment 1 (*i.e.*, participants *with* instruction on how to hold the device), three used first generation iPads, leaving only 98 participants in that group as well. As with Experiment 1, no lab-based participants were excluded as participant-level outliers.

We computed six measures from the accelerometer and gyroscope data: *extraneous movement* of the device, defined as the standard deviation of the accelerometer values in *x*, *y*, and *z* directions, and *device orientation*, based on the average gyroscope rotation around the *x*, *y*, and *z* axes. The device axes are shown in Figure 6. Gyroscope rotation around each axis ranges from -180° to 180°, but we take the absolute value of the rotation because we are interested in deviation from zero—for a perfectly horizontal orientation, rotations around the *x* and *y* axes are both zero. We also compare trial completion time and error rate, repeating a subset of the analysis from Experiment 1 but including the new participant group. For timing data, we again apply a log-transform and use a parametric test (one-way ANOVA), while we use separate non-parametric Kruskal-Wallis tests for all other measures to address normality issues. A Bonferroni adjustment was applied to all posthoc pairwise comparisons.

Results

Figures 7 and 8 show the extraneous touchscreen device movement and device rotation results.

Extraneous Device Movement

As shown in Figure 7, the average values for extraneous device movement were lowest with the lab-based group on all three axes, increasing for the crowdsourced group with instruction, and again for the crowdsourced group without instruction. Kruskal Wallis tests for each axis of movement showed that there were significant differences across the groups (*x* axis: $\chi^2_{df=2} = 43.11$, $p < .001$; *y* axis: $\chi^2_{df=2} = 50.65$, $p < .001$; z axis: $\chi^2_{df=2} = 62.99$, $p < .001$). Posthoc pairwise comparisons for each of the three axes using Mann Whitney U tests showed that the lab-based group had significantly lower extraneous device movement than both of the crowdsourced groups on all axes (all $p < .05$), and that the crowdsourced group with instruction had significantly lower movement on all three axes than the group without instruction (all $p < .05$). In other words, extraneous device movement increased as control decreased in terms of the environment, the instructions, or both.

Device Orientation

If the iPad is placed on a horizontal surface, orientation around the *x* and *y* axes should theoretically be zero, while *z* would vary. We thus focus on the *x* and *y* axes. As shown in Figure 8, rotation around these two axes increased as experimental control decreased. Kruskal Wallis tests for each axis of rotation showed that these differences were significant across the groups for both the *x* axis ($\chi^2_{df=2} = 77.23$, $p < .001$) and the *y* axis ($\chi^2_{df=2} = 43.98$, $p < .001$).

The angle of rotation was most pronounced around the *x*-axis—that is, tilting the top of the device closer to or farther from the user—where rotation varied from 1.21° on average (*SD* = 0.16) for the lab-based group to 32.26° (*SD* = 22.24) for the crowdsourced group without instruction. Rotation around the *y*-axis (tilting the device to the left or right) ranged from 0.58° on average (*SD* = 0.32) for the lab-based group to 5.99° (*SD* = 13.18) for the crowdsourced group without instruction. Posthoc pairwise comparisons using Mann Whitney U tests showed that both crowdsourced groups rotated the device more than the lab-based group along the *x* and *y* axes, and that the crowdsourced group without instruction did so more than the group with instruction (all comparisons $p < .05$).

For the *z* axis, there were no significant differences in rotation, as expected. Average rotations were 156.31° (*SD* = 72.38) for the lab-based group, 173.99° (*SD* = 86.81) for the crowdsourced group with instruction, and 170.92° (*SD* = 91.65) for the crowdsourced group without instruction.

Finally, of the 98 crowdsourced participants who did not receive instruction on how to hold/place the device, only 12 placed it approximately flat (absolute rotation of less than

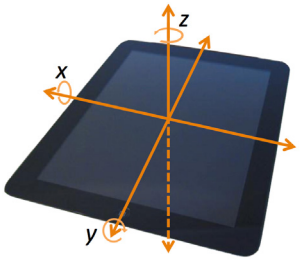

Figure 6. Device axes used for accelerometer and gyroscope analysis. If device is horizontal, rotations around *x* and *y* should theoretically be zero.

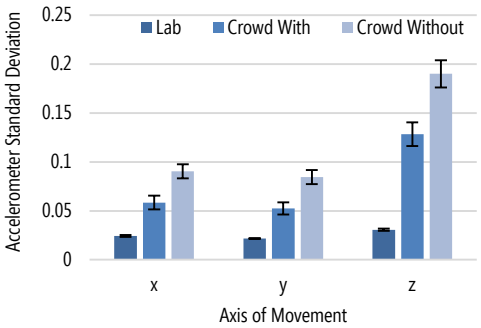

Figure 7. Extraneous device movement, increasing along all axes as experimental control decreases from lab to crowdsourced with and without instruction. (*N*=226 in total; error bars show std error.)

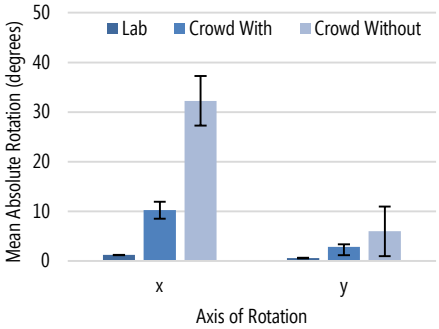

Figure 8. Rotation of the tablet, which increased as experimental control decreased: from lab to crowd to no placement instruction. The *z*-axis is not shown because all *z*-axis rotations are equally valid. (*N*=226 in total; error bars show std error.)

5° around both the *x* and *y* axes). In contrast, 62 of the crowdsourced participants with instruction placed the device flat—although this means that more than a third of this group still deviated from the study protocol. Lab participants all placed the device flat.

Trial Completion Time and Error Rate

Although two of the three participant groups here were included in Experiment 1's timing and error rate analyses, we conduct a secondary analysis to assess how that original data compares to the new crowdsourced group without instruction. For trial time and accuracy, the crowdsourced group without instruction was similar to that with instruction: average trial time of 765ms (SD = 215) and error rate of 6.8% (SD = 4.9) compared to 804ms (SD = 220) and 6.4% (SD = 5.0) for crowd participants with instruction. The lab group's average trial time was 832ms (SD = 185) and error rate was 3.0% (SD = 2.0).

These differences translated to a significant effect of participant group on error rate (Kruskal Wallis: $(\chi^2_{df=2} = 24.74, p < .001)$ but not on trial time (one-way ANOVA: $F_{2,223} = 1.78$, $p > .05$, $\eta_p^2 = .02$). Posthoc pairwise comparisons for error rate showed that the lab-based participants made fewer errors than both of the crowdsourced groups ($p < .05$), but there was no significant difference between the two crowdsourced groups.

Summary

As experimental control decreased—from the lab setting to the remote setting to removing the device placement instruction—both extraneous device movement and rotation of the device significantly increased. While our study design does not allow us to draw conclusions about causality, the result suggests that the decreased touchscreen input accuracy of crowdsourced participants in Experiment 1 may be related to the difficulty of controlling posture during mobile device testing outside of the lab setting. Interestingly, however, there were no timing or error rate differences between the two crowdsourced groups, suggesting that the differences between lab-based and crowdsourced participation is much larger than the impacts of tweaking instructions for crowdworkers.

DISCUSSION

Our findings show that collecting input performance data in the lab can differ from using a paid online crowdsourcing service such as MTurk. These differences were evident both in the conclusions reached by analyzing the lab-based and crowdsourced data separately and in statistically significant differences from a direct comparison between the two groups. In particular, the crowdsourced participants were more error prone with the touchscreen than the mouse, a finding that was opposite to the lab-based participants. An age-matched subset of crowdsourced participants also made a different speed-accuracy tradeoff than the lab-based participants, being faster and more error prone. Finally, extraneous mobile device movement and rotation of the device away from a horizontal position increased as experimental control decreased—from the lab to online to removal of instructions on device placement—and is a potential cause of the different conclusions regarding mobile input accuracy between the lab and the crowd.

These findings contrast those from Komarov *et al.*'s [21] crowdsourcing versus lab study with desktop interaction techniques. Indeed, their study found no "significant or substantial differences in the data collected in the two settings." The contrast is likely due both to the inclusion of the mobile device and to the larger sample size in our study (*i.e.*, 30 in-person participants compared to only 10–14 per experiment in [21]). Another difference, however, is in the tasks used. Komarov *et al.*'s bubble cursor experiment is most similar to our tasks, while their use of split menus and split interfaces could have incurred greater cognitive overhead and thus impacted results differently.

Considerations for Crowdsourced *vs.* Lab-based HCI Performance Studies

Our study shows that researchers cannot expect that crowdsourced data collection for performance experiments will directly reflect what would have been found in a more controlled setting, particularly for mobile input but also for desktop interaction techniques. The question, then, is which data source to use and when? While the best choice will ultimately depend on the individual study and its goals, the

following are important considerations based on both our findings and our experiences in conducting this research.

Ecological validity. Crowdsourced data is most useful if ecological validity is of high importance. Participants used their own computer setup for mouse input, different sizes of touchscreen tablets, and were more likely to hold their mobile device with a comfortable posture. Highlighting this lattermost point, only 12 of 98 participants in Experiment 2 placed their device flat on the table when given no instructions on device placement. Lab-based participants can also be instructed to hold the mobile device in whatever way is comfortable, but even this instruction would still not capture the range of body postures that occur in a real setting (*e.g.*, reclining on a couch vs. sitting at a table).

Experimental control. The study testbed included several features to attain experimental control for remote participants, such as providing redundant textual and visual instructions for all tasks and for device placement, and detecting long periods of inactivity during the task to restart the affected trial afterward. Despite these measures, differences in speed-accuracy tradeoffs and in evidence of following the mobile device placement instruction were seen between the crowd and the lab.

Participant incentives. Lab-based participants may feel a strong social bias to please the experimenter and follow instructions (*i.e.*, a stronger observer effect), while online participation may feel more casual, resulting in instructions not being taken as seriously. For crowdsourced participants, there is also a strong financial incentive to complete tasks quickly, which could impact speed-accuracy tradeoffs; we saw evidence of this impact in the comparison of age-matched participants. While it is not possible to determine from our study which data source ultimately better matches real-world behavior, these differences should be considered when interpreting results from either source and caution should be taken in comparing results across sources.

Cost. While running individual crowdsourced participants is cheaper than lab-based participants (especially given the experimenter's salary for lab sessions), a well-designed and vetted online testbed takes a substantial amount of time to build. In our case, this development cost was reasonable because of the scale of data we wanted to collect; the data reported here is part of a larger study. Another advantage is that, once built, it is trivial to collect more crowdsourced data (*e.g.*, for the verification reported in the Procedure section comparing small- vs. medium-sized iPads).

Overall, if a large amount of data is needed or ecological validity is of high importance, we recommend a crowdsourced approach despite the behavior differences seen between the crowd and the lab. For many smaller studies, however, a traditional lab experiment may be preferable. If opting for a crowdsourced study, well-specified instructions with redundant text and visual information (*e.g.*, Figure 5) can aid in imposing experimental control. We also suggest testing the online software thoroughly before deployment; indeed, we iteratively refined our instructions through pilot testing with tens of in-person and remote participants before the final deployment. Finally, when *interpreting* results from crowdsourced versus lab-based performance experiments, researchers should keep in mind potential differences in speed-accuracy tradeoffs and the degree to which mobile accuracy data, in particular, is valid.

Limitations and Future Work

Our study has limitations and leaves open opportunities for future work. First, the findings will generalize best to *paid* crowdsourcing services, such as MTurk. Other forms of remote participation, such as when there is an active remote facilitator or when data is collected within the context of a broader task (*e.g.*, a mobile game [14,28]) may yield different results. Second, we did not collect data on what iPad size crowdsourced participants used, so we cannot directly analyze a potential effect of device size. However, the theoretical and empirical analysis presented in the Procedure section of Experiment 1 demonstrates that this is not likely a concern. Future similar studies, of course, should collect this data for completeness. Third, our data did not allow us to understand the exact cause of the extra mobile device movement seen from the crowdsourced participants. For example, some participants may have placed the iPad on a flat but soft surface like their lap, which would have resulted in more movement than a tabletop. It could be useful to collect this data through self-report from remote participants or, alternatively, to experimentally control for a variety of common postures in a lab setting. Fourth, our study only included a basic set of input tasks. While the findings should largely translate to other desktop and mobile interaction techniques where measuring speed and accuracy is of utmost importance, it will be important to replicate the findings with more complex tasks that require more cognitive overhead.

CONCLUSION

We reported on two experiments to compare input performance data collected in person to that collected through a paid crowdsourcing service such as MTurk. Our findings show that study conclusions differ when analyzing data from the two sources separately—in particular, error rates when using a mobile touchscreen device were significantly higher than using a mouse for crowdsourced participants, whereas the opposite was true for lab-based participants. A potential contributing factor is that, with the mobile device, extraneous movement and variability in device orientation increased significantly as experimental control decreased, from the lab to the remote setting to removal of instructions on device placement. Crowdsourced participants were also significantly faster and more error prone than lab-based participants, perhaps due to differences in incentives. Overall, these tradeoffs should be considered by researchers who want to conduct HCI performance experiments using crowdsourcing methods.

ACKNOWLEDGMENTS

This project was funded by NSF grant IIS-1350438.

REFERENCES

1. Mohammad Allahbakhsh, Boualem Benatallah, A Ignjatovic, H R Motahari-Nezhad, E Bertino, and S Dustdar. 2013. Quality control in crowdsourcing systems. *IEEE Internet Computing* 17, 2: 76–81. http://doi.org/10.1109/MIC.2013.20
2. Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* 20, 3: 351–368. http://doi.org/10.1093/pan/mpr057
3. Xiaojun Bi, Yang Li, and Shumin Zhai. 2013. FFitts Law: Modeling Finger Touch with Fitts ' Law. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1363–1372. http://doi.org/10.1145/2470654.2466180
4. Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1: 3–5. http://doi.org/ 10.1177/1745691610393980
5. A. Cockburn, D. Ahlström, and C. Gutwin. 2012. Understanding performance in touch selections: Tap, drag and radial pointing drag with finger, stylus and mouse. *International Journal of Human-Computer Studies* 70, 3: 218–233. http://doi.org/10.1016/j.ijhcs.2011.11.002
6. Jay L Devore. 2015. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning.
7. Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 1013–1022. http://doi.org/10.1145/2145204.2145355
8. Carsten Eickhoff and Arjen P de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval* 16, 2: 121–137. http://doi.org/10.1007/s10791-011-9181-9
9. Leah Findlater, Jon E. Froehlich, Kays Fattal, Jacob O. Wobbrock, and Tanya Dastyar. 2013. Age-related differences in performance with touchscreens compared to traditional mouse input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, 343–346. http://doi.org/10.1145/2470654.2470703
10. Tovi Grossman and Ravin Balakrishnan. 2005. 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*, 281–290. http://doi.org/10.1145/1054972.1055012
11. Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203–212. http://doi.org/10.1145/1753326.1753357
12. Niels Henze, Enrico Rukzio, and Susanne Boll. 2011. 100,000,000 taps: analysis and improvement of touch performance in the large. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'11)*, 133–142. http://doi.org/10.1145/2037373.2037395
13. John J Horton, David G Rand, and Richard J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* 14, 3: 399–425. http://doi.org/10.1007/s10683-011-9273-9
14. International Organization for Standardization. 2002. *Ergonomic requirements for office work with visual display terminals (VDTs)—Requirements for non-keyboard input devices.*
15. Panagiotis G Ipeirotis. 2010. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students* 17, 2: 16–21. http://doi.org/10.1145/1869086.1869094
16. Panos Ipeirotis. 2016. MTurk Tracker. Retrieved August 28, 2016 from http://demographics.mturk-tracker.com/
17. Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 453–456. http://doi.org/10.1145/1357054.1357127
18. Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, et al. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 1301–1318. http://doi.org/10.1145/2441776.2441923
19. Steven Komarov, Katharina Reinecke, and Krzysztof Z Gajos. 2013. Crowdsourcing Performance Evaluations of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 207–216. http://doi.org/10.1145/2470654.2470684
20. I. Scott MacKenzie and Poika Isokoski. 2008. Fitts' throughput and the speed-accuracy tradeoff. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*, 1633–1636. http://doi.org/10.1145/1357054.1357308
21. Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1: 1–23. http://doi.org/10.3758/s13428-011-0124-6

22. Gabriele Paolacci and Jesse Chandler. 2014. Inside the turk understanding mechanical turk as a participant pool. *Current Directions in Psychological Science* 23, 3: 184–188. http://doi.org/10.1177/0963721414531598

23. Alexander J Quinn and Benjamin B Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*, 1403–1412. http://doi.org/10.1145/1978942.1979148

24. Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, 2863–2872. http://doi.org/10.1145/1753846.1753873

25. Dmitry Rudchenko, Tim Paek, and Eric Badger. 2011. Text Text Revolution: a game that improves text entry on mobile touchscreen keyboards. In *Proceedings of Pervasive 2011*, 206–213. http://doi.org/10.1007/978-3-642-21726-5_13

26. M. Six Silberman, Kristy Milland, Rochelle LaPlante, Joel Ross, and Lilly Irani. 2015. Stop citing Ross et al. 2010, "Who are the crowdworkers?" Retrieved August 28, 2016 from https://medium.com/@silberman/stop-citing-ross-et-al-2010-who-are-the-crowdworkers-b3b9b1e8d300#.hv8r2sjcy

27. R. William Soukoreff and I. Scott MacKenzie. 2004. Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *International Journal of Human-Computer Studies* 61, 6: 751–789. http://doi.org/10.1016/j.ijhcs.2004.09.001

28. Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 143–146. http://doi.org/10.1145/1978942.1978963

29. Jacob O Wobbrock, Kristen Shinohara, and Alex Jansen. 2011. The effects of task dimensionality, endpoint deviation, throughput calculation, and experiment design on pointing measures and models. *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*: 1639. http://doi.org/10.1145/1978942.1979181