Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2021)

Utility of Crowdsourced User Experiments for Measuring the Central Tendency of User Performance to Evaluate Error-Rate Models on GUIs

Shota Yamanaka
Yahoo Japan Corporation
Kioi-cho 1-3, Chiyoda-ku, Tokyo, Japan
syamanak@yahoo-corp.jp

Abstract

The usage of crowdsourcing to recruit numerous participants has been recognized as beneficial in the human-computer interaction (HCI) field, such as for designing user interfaces and validating user performance models. In this work, we investigate its effectiveness for evaluating an error-rate prediction model in target pointing tasks. In contrast to models for operational times, a clicking error (i.e., missing a target) occurs by chance at a certain probability, e.g., 5%. Therefore, in traditional laboratory-based experiments, a lot of repetitions are needed to measure the central tendency of error rates. We hypothesize that recruiting many workers would enable us to keep the number of repetitions per worker much smaller. We collected data from 384 workers and found that existing models on operational time and error rate showed good fits (both $R^2 > 0.95$). A simulation where we changed the number of participants N_P and the number of repetitions N_{repeat} showed that the time prediction model was robust against small N_P and N_{repeat}, although the error-rate model fitness was considerably degraded. These findings empirically demonstrate a new utility of crowdsourced user experiments for collecting numerous participants, which should be of great use to HCI researchers for their evaluation studies.

Introduction

In the field of human-computer interaction (HCI), a major topic is to measure the time needed to complete a given task. Examples include measuring a text-entry time (Banovic et al. 2019; Cui et al. 2020), a time to learn a new keyboard layout (Jokinen et al. 2017), and a menu-selection time (Bailly, Lecolinet, and Nigay 2016). In these studies, generally, laboratory-based user experiments have been conducted. That is, researchers recruit 10–20 students from a local university and ask them to use a specified apparatus to perform a task in a silent room. However, researchers are recently aware of the risk of conducting a user experiment with a small sample size; e.g., the statistical power is weak (Caine 2016).

For research involving user experiments on graphical user interfaces (GUIs), it has recently become more common to recruit workers through crowdsourcing services (Cockburn et al. 2020; Findlater et al. 2017; Komarov, Reinecke, and Gajos 2013; Matejka et al. 2016; Yamanaka, Shimono, and Miyashita 2019). Previous studies have typically focused on designing better GUIs or conducting user experiments to evaluate novel interaction techniques compared with baselines to demonstrate that a proposed method is statistically better than a baseline.

Another major topic in HCI is deriving user performance models and empirically validating them. Conventionally, there are two representative metrics for GUI operations to be modeled: time and error rate (Wobbrock et al. 2008). A well-known model to predict the operational time for target pointing tasks is Fitts' law (Fitts 1954), or referred to as *Fitts's law* in some papers (MacKenzie 2002). In user experiments to evaluate the model fitness in terms of R^2, ten or 20 university student participants typically join a study and are asked to point to a target repeatedly. For example, researchers set three target distances and three target sizes (i.e., nine task conditions in total), and the participants repeatedly click a target 15 times for each task condition. The average time for these 15 clicks is recorded as the final score for a participant (Soukoreff and MacKenzie 2004).

In addition to operation times, the importance of predicting how accurately users can perform a task has recently been emphasized (Bi and Zhai 2016; Do, Chang, and Lee 2021; Huang et al. 2018, 2020; Park and Lee 2018; Yamanaka et al. 2020). In contrast to measuring the target-pointing times, where the time to click a target can be measured in every trial, the error rate is computed after repeatedly performing a single task condition (15 trials in the above-mentioned case). For example, if a participant misses a target in one trial, the error rate is recorded as $1/15 \times 100 = 6.67\%$; if there are ten participants, one miss corresponds to 0.667% in the end. Because errors can occur by chance, evaluating error-rate models often requires more data (repetitions) for each task condition to measure the central tendency of the error rate. To evaluate the model's prediction accuracy more precisely, researchers have asked participants to perform more repetitions, as it is often difficult to collect numerous participants for lab-based experiments. For example, a previous study on touch-based error-rate models set 40 repetitions for each task condition collected from 12 participants. In this case, one miss corresponded to a 0.208% error rate (Yamanaka and Usuba 2020).

However, for crowdsourced user experiments with GUIs,

researchers cannot set a large number of repetitions per task condition. To enable crowdworkers to concentrate on a given task, it is recommended to set short task completion times, as workers switch to other tasks every 5 min on average (Gould, Cox, and Brumby 2016). Hence, forcing a routine GUI operation task that takes, e.g., 40 min (Huang et al. 2018) or 1 hour (Park and Lee 2018; Yamanaka et al. 2020) would be harmful in terms of accurate measurement of the error rates. This could be considered a disadvantage of crowdsourced GUI study. An alternative to increasing the number of repetitions per task condition is simply to recruit more workers. This would enable the error rates to be measured more precisely, which would lead to a good prediction accuracy by the error-rate model (**our research hypothesis**). Even if the number of repetitions is only ten, utilizing 300 workers would mean that one miss corresponds to 0.033%. This is much more precise than the above-mentioned examples with error rates such as 0.208%.

However, there are several crowdsourcing-specific uncertainties that might affect the user performance results. For example, crowdworkers use different mice, displays, operating systems, cursor speed configurations, and so on; these factors significantly affect the target pointing performance in terms of both time and accuracy (Casiez and Roussel 2011; MacKenzie, Kauppinen, and Silfverberg 2001). In addition, while studies have shown that the performance model on time (Fitts' law) is valid for crowdsourced data, crowdworkers tend to be more inaccurate than lab-based participants in target pointing tasks (Komarov, Reinecke, and Gajos 2013), where error rates approximately two times higher or more have been observed (Findlater et al. 2017). Therefore, we would avoid claiming that user-performance models validated in crowdsourced studies are always applicable to lab-based controlled experiments. Also, it is not reasonable to interpret that the results such as error rates and operational times are directly comparable with lab-based participants; rather, we should consider that our results are *internally* valid.

Nevertheless, if an error-rate model we test exhibits a good fit (e.g., $R^2 > 0.9$), HCI researchers would have access to a powerful tool, crowdsourcing, to evaluate their newly proposed error-rate prediction models. Such a result stands to expand the application range of crowdsourcing in HCI; this motivated us to conduct this work. Our contributions are as follows.

- We conducted a crowdsourced mouse-pointing experiment following the Fitts' law paradigm. In total, we recorded 92,160 clicks performed by 384 crowd workers. Our error-rate model showed a good fit with $R^2 = 0.9581$, and cross-validation confirmed that the model can predict new (unknown) task conditions, too. This is the first study that demonstrates a GUI error-rate model holding to crowdsourced user data.
- We simulated how the number of participants N_P and the number of repetitions per task condition N_{repeat} affected the model fitness. We randomly sampled a limited portion of the entire workers (N_P from 10 to 320), and while each worker performed ten trials per task condition, we used only the data for the first N_{repeat} trials (from 2 to 10). After testing the model fitness over 100 iterations, we found that increasing N_P improved the prediction accuracy as well as increasing N_{repeat} could. The effect of N_P and N_{repeat} on the fitness was more clearly observed for the error-rate model than the time model, which suggests that crowdsourcing services are more suitable for evaluating novel error-rate models.

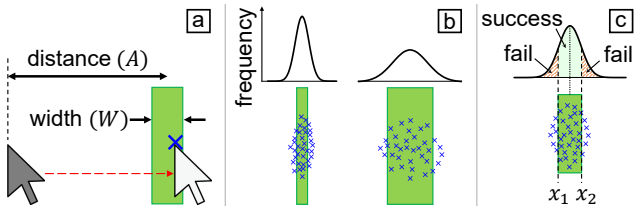

Figure 1: (a) We use the Fitts' law paradigm in which users point to a vertically long target. A clicked position is illustrated with an 'x' mark. (b) It has been assumed that the click positions recorded in many trials distribute normally, and its variability would increase with the target width. (c) An error rate is computed based on the probability where a click falls outside the target.

Related Work

Time Prediction for Pointing Tasks

For comparing the sensitivity of time and error-rate prediction models against N_P and N_{repeat}, we examine a robust time-prediction model, called Fitts' law (Fitts 1954). According to this model, the time for the first click, or movement time MT, to point to a target is related to the index of difficulty ID measured in bits:

$$MT = a + b \cdot ID = a + b \cdot (A/W + 1), \qquad (1)$$

where a and b are empirical regression constants, A is the target distance (or amplitude), and W is its width (see Figure 1a). There are numerous formulae for calculating the ID, such as using a square root instead of the logarithm or using the effective target width (Plamondon and Alimi 1997), but previous studies have shown that Equation 1 yields excellent model fitness (Soukoreff and MacKenzie 2004). Using this Fitts' law, researchers can measure MTs for several $\{A, W\}$ conditions, regress the data to compute a and b, and then predict the MT for a new $\{A, W\}$ condition by applying the parameters of $\{a, b, A, W\}$ to Equation 1.

Error-rate Prediction for Pointing Tasks

Researchers have also tried to derive models to predict the error rate ER (Meyer et al. 1988; Wobbrock et al. 2008; Park and Lee 2018). In practice, the ER should increase as participants move faster, and vice versa (Batmaz and Stuerzlinger 2021; Zhai, Kong, and Ren 2004). In typical target pointing experiments, participants are instructed to "point to the target as quickly and accurately as possible," which is intended to balance the speed and carefulness to decrease both MT and ER (MacKenzie 1992; Soukoreff and MacKenzie 2004).

In pointing tasks, as the target size decreases, users have to aim for the target more carefully to avoid misses. Accordingly, the spread of click positions should be smaller. If researchers conduct a pointing experiment following a typical Fitts' law methodology, in which two vertically long targets are used and participants perform left-right cursor movements, the click positions would follow a normal distribution (Figure 1b) (Crossman 1956; MacKenzie 1992). Formally speaking, a click point is a random variable X following normal distribution: $X \sim \mathcal{N}(\mu, \sigma^2)$, where μ and σ are the mean and standard deviation of the click positions on the x-axis, respectively. The click point variability σ is assumed to proportionally relate to the target width, or to need an intercept, i.e., linear relationship (Bi and Zhai 2016; Yamanaka and Usuba 2020; Yu et al. 2019):

$$\sigma = c + d \cdot W, \tag{2}$$

where c and d are regression constants. The probability density function for a normal distribution, $f(x)$, is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{3}$$

If we define the target center as located at $x = 0$ with the target boundary ranging from x_1 to x_2 (Figure 1c), the predicted probability for where the click point X falls on the target, $P(x_1 \leq X \leq x_2)$, is

$$\int_{x_1}^{x_2} f(x)dx = \frac{1}{2}\left[\mathrm{erf}\left(\frac{x_2-\mu}{\sigma\sqrt{2}}\right) - \mathrm{erf}\left(\frac{x_1-\mu}{\sigma\sqrt{2}}\right)\right], \tag{4}$$

where $\mathrm{erf}(\cdot)$ is the Gauss error function:

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt. \tag{5}$$

Previous studies have shown that the mean click point is located close to the target center ($\mu \approx 0$), and μ is not significantly affected by the target distance A (Bi and Zhai 2016; MacKenzie 1992; Yamanaka and Usuba 2020). Given the target width W, Equation 4 can be simplified and the ER is predicted as

$$ER = 1 - P\left(-\frac{W}{2} \leq X \leq \frac{W}{2}\right) = 1 - \mathrm{erf}\left(\frac{W}{2\sqrt{2}\sigma}\right). \tag{6}$$

Similarly to the way Fitts' law is used, researchers measure σ for several $\{A, W\}$ conditions, regress the data to compute c and d in Equation 2, and then predict the σ for a new $\{A, W\}$ condition. In this way (i.e., using the predicted σ based on a new W), we can predict the ER with Equation 6 for a new task condition. While there are similar but more complicated versions of this model tuned for pointing tasks in virtual reality systems (Yu et al. 2019) and touchscreens (Bi and Zhai 2016), to our knowledge, there has been no report on the evaluation of this model for the most fundamental computer environment, i.e., PCs with mice.

Crowdsourced Studies on User Performance and Model Evaluation for GUIs

For target pointing tasks in PC environments, Komarov et al. found that crowdsourced and lab-based experiments led to the same conclusions on user performance, such as that a novel facilitation technique called *Bubble Cursor* (Grossman and Balakrishnan 2005) reduced the MT compared with the baseline point-and-click method (Komarov, Reinecke, and Gajos 2013). Yamanaka et al. tested the effects of target margins on touch-pointing performance using smartphones and reported that the same effects were consistently found in crowdsourced and lab-based experiments, e.g., wider margins significantly decreased the MT but increased the ER (Yamanaka, Shimono, and Miyashita 2019). Findlater et al. showed that crowdworkers had significantly shorter MTs and higher ERs than lab-based participants in both mouse- and touch-pointing tasks (Findlater et al. 2017). Thus, they concluded that crowdworkers were more biased towards speed than accuracy when instructed to "operate as quickly and accurately as possible."

Regarding Fitts' law fitness, Findlater et al. reported that crowdworkers had average values of Pearson's $r = 0.926$ with mice and $r = 0.898$ with touchscreens (Findlater et al. 2017). Schwab et al. conducted crowdsourced scrolling tasks and found that Fitts' law held with $R^2 = 0.983$ and 0.972 for the desktop and mobile cases, respectively (Schwab et al. 2019) (note that scrolling operations follow Fitts' law well (Zhao et al. 2014)). Overall, these reports suggest that Fitts' law is valid for crowdsourced data regardless of the input device. It is unclear, however, how the N_P affects model fitness, because these studies used the entire workers' data for model fitting.

The only paper that tested the effect of N_P on the fitness of user-performance models is a recent work by Yamanaka (Yamanaka 2021). He tested modified versions of Fitts' law to predict MTs in a rectangular-target pointing task. The conclusion was that, although he changed N_P from 5 to 100, the best-fit model did not change. However, because he used all N_{repeat} clicks, increasing N_P always increased the total data points to be analyzed, and thus the contributions of N_P and N_{repeat} could not be analyzed separately. We further analyze this point in our simulation.

In summary, there is a consensus that a time prediction model for pointing tasks (Fitts' law) shows a good fit for crowdsourced data. However, ER data have typically been reported as secondary results when measuring user performance in these studies. At least, no studies on evaluating ER prediction models have been reported so far. If we can demonstrate the potential of crowdsourced ER model evaluation, at least for one example task (target pointing in a PC environment), it will motivate future researchers to investigate novel ER models with less recruitment effort, more diversity of participants, and less time-consuming data collection. This will directly benefit the contribution of crowdsourcing to the HCI field.

User Experiment

We conducted a traditional cyclic target-pointing experiment on the *Yahoo! Crowdsourcing* platform (https://crowdsourcing.yahoo.co.jp). Our affiliation's IRB-equivalent research ethics team approved this study. The experimental system was developed with the `Hot`

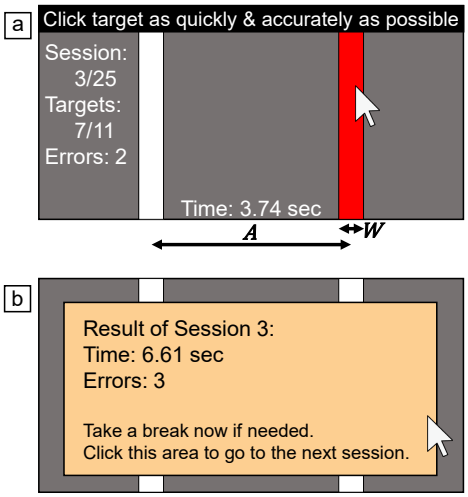

Figure 2: Task stimuli used in the experiment. (a) Participants clicked alternately on each target when it was red. (b) At the end of a session, the results and a message to take a break were shown.

`Soup Processor` programming language. The crowdworkers were asked to download and run an executable file to perform the experimental task.

Task, Design, and Procedure

In the task window (1200 × 700 pixels), two vertically long targets were displayed (Figure 2a). If the participants clicked the target, the red target and white non-target rectangles switched colors, and they successively performed this action back and forth. If the participants missed the target, it flashed yellow, and they had to keep trying until successfully clicking it. We did not give auditory feedback for success or failure, as not all the participants would have been able to hear sound during the task. A *session* consisted of 11 cyclic clicks with a fixed $A \times W$ condition. After completing a session, the participant saw the results and a message to take a break (Figure 2b).

The experiment was a 3×8 within-subjects repeated-measures design with the following independent variables and levels: three target distances (A = 300, 460, and 630 pixels) and eight widths (W = 8, 12, 18, 26, 36, 48, 62, and 78 pixels). These values were selected so that the values of ID ranged widely from 2.28 to 6.32 bits, which sufficiently covered easy to hard conditions according to a survey (Soukoreff and MacKenzie 2004). Each participant completed 24 ($= 3_A \times 8_W$) sessions. The order of the 24 conditions was randomized. Before the first session, to allow the participants to get used to the task, they performed a practice session under a condition with $A = 400$ and $W = 31$ pixels, i.e., parameters that were not used in the actual 24 data-collection sessions. This experimental design was tuned with reference to the authors' pilot study; without having a break, the task completion time was 3 min 40 sec on average, which meets the recommendation for crowdsourced user experiments (Gould, Cox, and Brumby 2016).

The MT was measured from when the previous target was successfully clicked to when the next click was performed regardless of the success or failure (MacKenzie 1992; Soukoreff and MacKenzie 2004). Trials in which we observed one or more clicks outside the target were flagged as an error. The first left target acted as a starting button, and the remaining ten trials' data were measured to compute MT, σ, and ER. After finishing all sessions, the participants completed a questionnaire on their age (numeric), gender (free-form to allow non-binary or arbitrary answers), handedness (left or right), Windows version (free-form), input device (free-form), and history of PC use (numeric in years).

Participants and Recruitment

We recruited workers who used Windows Vista or a later version to run our system. We requested no specific PC skills, as we did not wish to limit our collection to only high-performance workers' data. Also, we did not use any *a-priori* filtering options, such as the approval-rate threshold, which require additional cost for the crowdsourcing service. We made this decision because, if our hypothesis is supported with a less costly method, it would be more beneficial for future research to recruit many more participants with low cost for obtaining the central tendency of error rates. Still, clear outlier workers who seemed not to follow our instructions (such as performing the task too slowly) were removed when we analyzed the data. As we show later in the simulation analysis, this decision was not problematic because Fitts' law held well even if we analyzed only ten workers' data over 100 iterations (i.e., they exhibited typical rapid-and-accurate pointing behavior).

On the recruitment page, we asked the workers to use a mouse if possible. We made this request because, in our simulation analysis, we randomly selected a certain number of participants (e.g., $N_P = 10$) to examine if the model fitness was good or poor. If these workers used different devices (e.g., six mice, two touchpads, and two trackballs), we might have wondered if a poor model fit was due to the device differences. Nevertheless, to avoid a possible false report in which all workers might answer they used mice, we explicitly explained that any device was acceptable, and then removed the non-mouse users from the analysis.

Once workers accepted the task, they were asked to read the online instructions, which stated that they should perform the task as rapidly and accurately as possible. This was also always written at the top of the experimental window as a reminder (Figure 2a). After they finished all 25 sessions and completed the questionnaire, the log data was exported to a csv file. They uploaded the file to a server and then received a payment of JPY 100 (∼USD 0.92).

In total, 398 workers completed the task, including 384 mouse users. The mouse users' demographics were as follows. Age: 16 to 76 years, with $M = 43.6$ and $SD = 11.0$. Gender: 300 male, 79 female, and 5 chose not to answer. Handedness: 24 were left-handed and 360 were right-handed. Windows version: 1 used Vista, 27 used Win7, 8 used Win8, and 348 used Win10. PC usage history: 0 (less than 1 year) to 45 years, with $M = 21.8$ and $SD = 7.82$.

In this study, we do not analyze these demographic data in detail. For example, it has been reported that participants' handedness (Hoffmann 1997), gender and age (Brogmus 1991) affect Fitts' law performance. In our simulation, it is possible that the data may be biased; e.g., when we select $N_P = 10$ workers, they are all males in their 60s. If researchers want to investigate this point, controlling the sampled workers' demographics before executing the simulation is needed.

For mouse users, the main pointing task took 3 min 45 sec on average without breaks. With breaks, the mean task completion time was 5 min 42 sec, and thus the effective hourly payment was JPY 1,053 (∼USD 9.69). Note that this effective payment could change depending on other factors such as the times for reading the instructions and for uploading the csv file.

Results

Outlier Data Screening

Following previous studies (Findlater et al. 2017; MacKenzie and Isokoski 2008), we removed trial-level spatial outliers if the distance of the first click position was shorter than half of target distance $A/2$ (i.e., clicking closer to the non-target than the target) to omit clear accidental operations such as double-clicking the previous target. Another criterion used in these studies was to remove trials in which the click position was more than twice of target width $2W$ away from the target center. We did not use this criterion, as we would like to measure error trials even where a click position was $\geq (2W + 1)$ pixels away from the target center.

To detect trial-level temporal outliers to remove extremely fast or slow operations, we used the inter-quartile range (IQR) method (Devore 2011), which is more robust than the *mean-and-3σ* approach. The IQR is defined as the difference between the third and first quartiles of the MT for each session for each participant. Trials in which the MT was more than $3 \times IQR$ higher than the third quartile or more than $3 \times IQR$ lower than the first quartile were removed.

For participant-level outliers, we calculated the mean MT across all 24 conditions ($3_A \times 8_W$) for each participant. Then, using each participant's mean MT, we again applied the IQR method and removed extremely rapid or slow participants. The trial- and participant-level outliers were independently detected and removed.

As a result, among the 92,160 trials ($= 3_A \times 8_W \times 10_{\text{repetitions}} \times 384_{\text{workers}}$), we identified 1,191 trial-level outliers (1.29%). We also found two participant-level outlier workers. While the mean MT of all participants was 898 ms and the IQR was 155 ms, the outlier workers' mean MTs were 1,462 and 1,533 ms. Accordingly, the data from all 480 trials were removed ($= 3_A \times 8_W \times 10_{\text{repetitions}} \times 2_{\text{participants}}$). They also exhibited trial-level outliers (i.e., there were overlaps); as a result, the data from 1,664 trials were removed in total (1.81%), which was close to the rate in a previous study (Findlater et al. 2017).

Analyses of Dependent Variables

After the outliers were removed, the data from 90,496 trials (98.2%) were analyzed. The dependent variables were the MT, σ, and ER.

Movement Time. We used the Shapiro-Wilk test ($\alpha = 0.05$) and Q-Q plot to check the normality assumption required for parametric ANOVAs. The MT data did not pass the normality test, and thus we log-transformed the data to meet the normality assumption. The log-transformed data passed the normality test, and we used RM-ANOVAs with Bonferroni's p-value adjustment method for pairwise comparisons. For the F statistic, the degrees of freedom were corrected using the Greenhouse-Geisser method when Mauchly's sphericity assumption was violated ($\alpha = 0.05$).

We found significant main effects of A ($F_{1.909,727.1} = 2674$, $p < 0.001$, $\eta_p^2 = 0.88$) and W ($F_{4.185,1595} = 6813$, $p < 0.001$, $\eta_p^2 = 0.95$) on MT. A significant interaction was found for $A \times W$ ($F_{13.01,4955} = 14.23$, $p < 0.001$, $\eta_p^2 = 0.036$). As shown in Figure 3a, the MT increased as the ID increased (i.e., A increased or W decreased). Fitts' law showed an excellent fitness with $R^2 = 0.9789$.

Click Point Variability. The σ data and its log-transformed data did not pass the normality test, and thus we used a non-parametric ANOVAs with aligned rank transform (Wobbrock et al. 2011) with Tukey's p-value adjustment method for pairwise tests. We found significant main effects of A ($F_{2,762} = 3.683$, $p < 0.05$, $\eta_p^2 = 0.0096$) and W ($F_{7,2667} = 6043$, $p < 0.001$, $\eta_p^2 = 0.94$) on σ. An interaction of $A \times W$ was not significant ($F_{14,5334} = 0.8411$, $p = 0.62$, $\eta_p^2 = 0.0022$). The model fitness of Equation 2 ($\sigma = a + b \cdot W$) was quite high ($R^2 = 0.9966$), as shown in Figure 3b.

Our model assumes that σ is not affected by A, but the result showed that A significantly affected σ. This statistical significance likely comes from the large number of participants. When we checked this in more detail, we found that the effect size of A was quite small compared with W ($\eta_p^2 = 0.0096$ vs. 0.94, respectively), and the mean σ values for $A = 300$, 460, and 630 pixels were 7.258, 7.293, and 7.309 pixels, which fall within a 0.051-pixel range ($< 1\%$). In contrast, the σ values varied from 2.17 to 14.2 pixels due to W. While we plotted 24 points ($3_A \times 8_W$) in Figure 3b, it looks as though there were only eight points, as the three σ values for the three As were almost the same and thus they overlapped.

Error Rate. The ER data and its log-transformed data did not pass the normality test, and thus we again used a non-parametric ANOVAs with aligned rank transform. We found significant main effects of A ($F_{2,762} = 6.732$, $p < 0.01$, $\eta_p^2 = 0.017$) and W ($F_{7,2667} = 96.90$, $p < 0.001$, $\eta_p^2 = 0.20$) on ER. An interaction of $A \times W$ was not significant ($F_{14,5334} = 1.627$, $p = 0.064$, $\eta_p^2 = 0.0043$).

Using Equations 2 and 6, we can predict the ERs based on given W values. The predicted and actually observed ERs are shown in Figure 3c. The worst prediction error was 4.235 points in the case of $(A, W) = (300, 8)$. As a comparison,

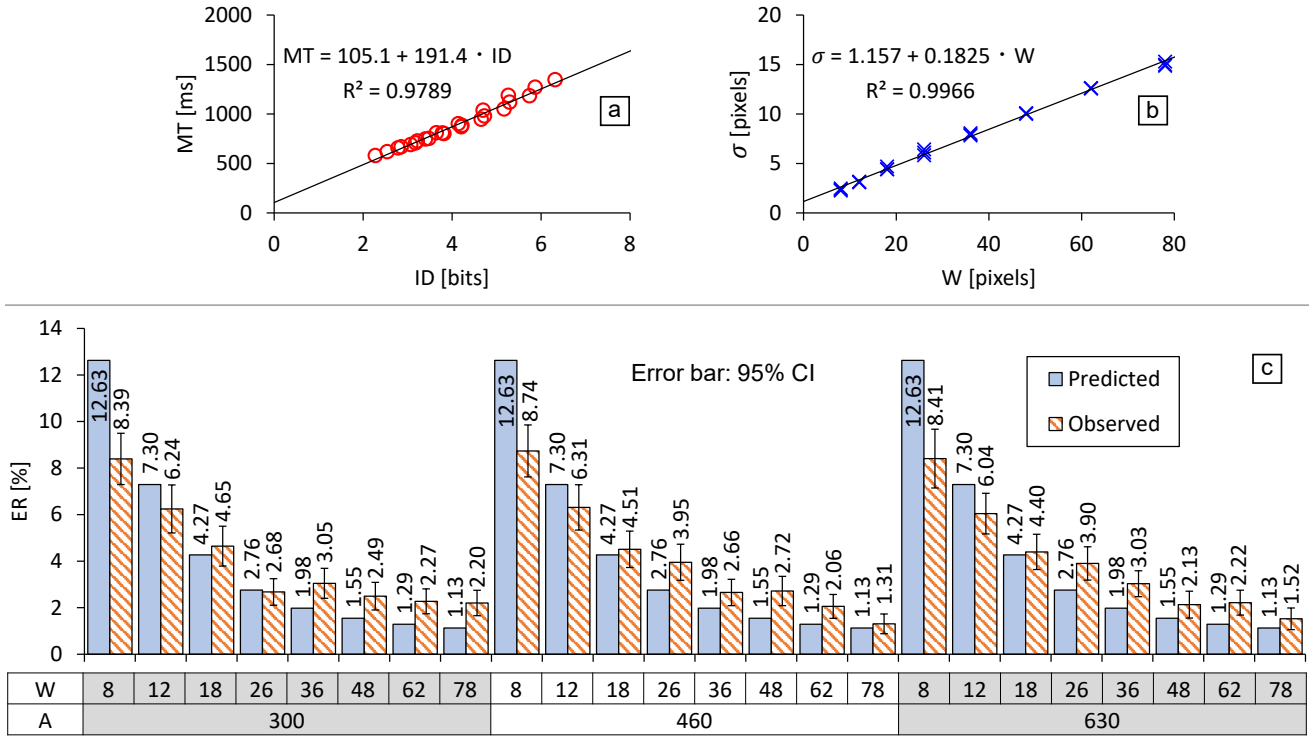

Figure 3: Model fitness results for (a) Fitts' law and (b) click point variability. (c) Comparison of the predicted vs. observed ERs.

previous studies on touch-based pointing tasks have reported that the prediction error for $W = 2.4$-mm targets was 9.74 points (Bi and Zhai 2016) and that for 2-mm was 10.07 points (Yamanaka and Usuba 2020). While a direct comparison with touch operations is not particularly fruitful, the tendency that prediction errors increase for smaller Ws is consistent between the previous studies and ours.

To formally evaluate our model's prediction accuracy, we computed the following three fitness criteria. The correlation between predicted vs. observed ERs was $R^2 = 0.9581$. The mean absolute error MAE was 1.193%. The root mean square error $RMSE$ was 1.665%. In addition, to evaluate the prediction accuracy for new (unknown) task conditions, we ran a leave-one-(A, W)-out cross-validation. The three criteria for the ER prediction were $R^2 = 0.9529$, $MAE = 1.272\%$, and $RMSE = 1.814$. The worst prediction error was 4.805 points. These results indicate that, even for researchers who would like to predict the ER for a new task condition based on previously measured data, the prediction accuracy would not be considerably degraded.

Simulation

Although our N_{repeat} (10) was not large compared with previous studies on error-rate prediction models due to the time constraint for crowdsourcing, we hypothesized that increasing N_P would improve the model fitness. We also wonder how the model fitness changes when N_{repeat} is much smaller, which further shortens the task completion time for workers. For example, if it were 5, the average task completion time would be 2 min 51 sec including breaks (i.e., half of 5 min 42 sec). Note that N_{repeat} must be greater than 1 to compute the standard deviation σ.

We randomly selected N_P workers' data from the 384 mouse users by changing N_P from 10 (typical lab-based experiments) to 320 by doubling it repeatedly. The N_{repeat} changed from 2 to 10; if it was 2, we used only the first two repetitions' data and the subsequent eight trials were removed. Outlier detection was run in the same manner as if we had conducted an experiment newly with N_P workers. Then, we analyzed the R^2 values for Equations 1 (Fitts' law), 2 (click point variability σ), and 6 (ER). To handle the randomness to select N_P workers, we ran this process over 100 iterations and averaged the R^2 values for each of $N_P \times N_{\text{repeat}}$.

The results are shown in Figure 4. First, we can visually confirm that the time prediction model (a) showed the flattest fitness compared with the other two models (c) and (e). The R^2 values were consistently over 0.90, and after we collected 20 participants or measured four repetitions, R^2 was over 0.95 (b). This result supports the decision of previous studies' lab-based experiments that recruited ten to 20 participants to examine Fitts' law. While repeating 15 to 25 trials per task condition has been recommended (Soukoreff and MacKenzie 2004), our results show that a much smaller number of repetitions will suffice.

For the click point variability, as (c) shows, the model fitness was relatively worse only when both N_P and N_{repeat} are small. The increase in either N_P or N_{repeat} can resolve

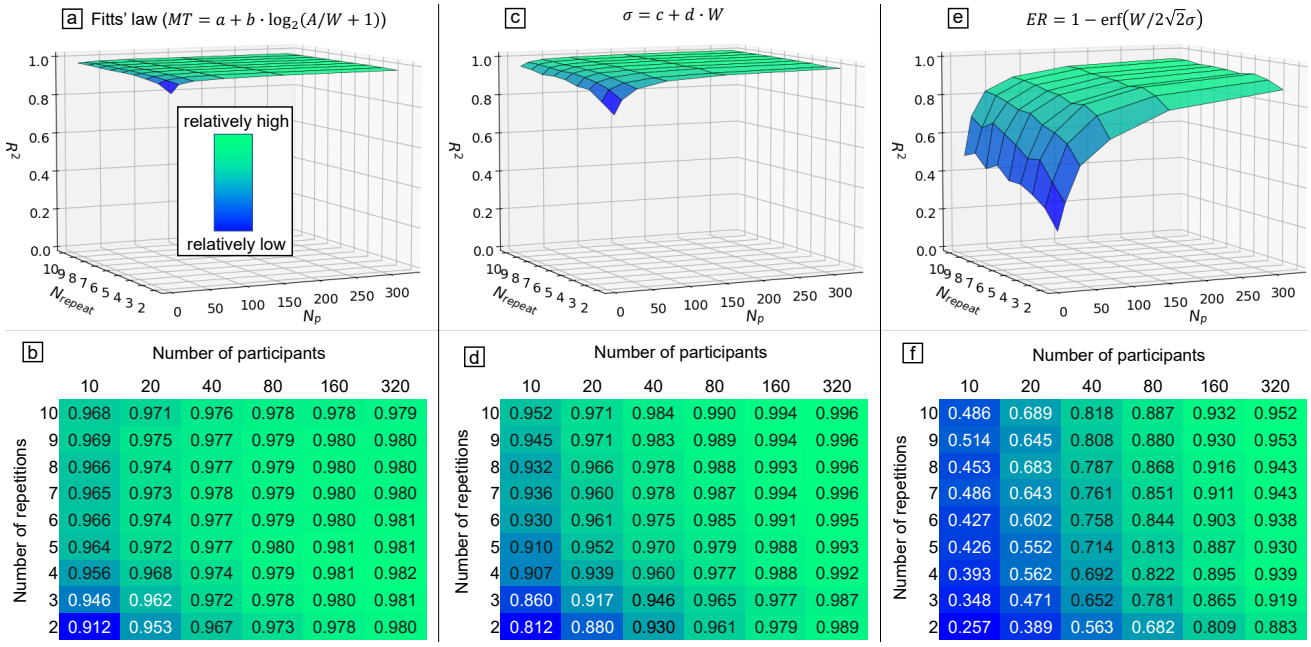

Figure 4: Simulation results on mean model fitness in R^2 by changing N_P and N_{repeat} over 100 iterations.

this. For example, by collecting $N_P \geq 80$ workers or repeating ten trials, we obtain $R^2 > 0.95$.

Lastly, for the error-rate model, the fitness was affected by N_P and N_{repeat} most drastically, as shown in (e). Particularly for small N_P values such as 10 and 20, the R^2 values were less than 0.70 (f), which is a unique result compared with the other two models that always showed R^2 greater than 0.80 in (b) and (d). If we fully use ten repetitions and would like to obtain a certain value of the model fitness (such as $R^2 > 0.9$), collecting 160 participants is sufficient—more precisely, when we tested N_P from 80 to 160 (step: 1), $N_P = 92$ achieved $R^2 = 0.9029 > 0.9$ for the first time.

Figure 4e–f demonstrates that increasing N_P can be a viable alternative to increasing N_{repeat} to obtain a higher prediction accuracy for this error-rate model. Suppose we have a case where researchers want to set a smaller N_{repeat} such as 3 instead of 10 due to (e.g.) asking workers to answer more questionnaire items after the task. Even for this case, by collecting $N_P = 320$ workers, the model would fit to the data with $R^2 > 0.9$ in our data. Hence, although the task completion time for crowdsourced user experiments should not be too long (Gould, Cox, and Brumby 2016), the easy recruitment for crowdsourcing enables researchers to measure the central tendency of error rates. This benefit of crowdsourcing is more critical for error-rate models than time-prediction models, as we demonstrated here, which has never been empirically reported before.

Discussion

Benefits and Implications of Using Crowdsourcing for Error-rate Model Evaluation

In this study, we explored the potential of crowdsourcing for evaluating error-rate prediction models on GUIs. As one of the most fundamental operations, we utilized a Fitts' law task for its well-structured methodology. The results obtained from 384 crowdworkers showed that the models on Fitts' law and the click point variability fit well to the empirical data with $R^2 = 0.9789$ and 0.9966, respectively, as shown in Figure 3a–b. Using the predicted σ values based on W, we then predicted the ERs for each $A \times W$ condition, which yielded the correlation between predicted vs. observed ERs of $R^2 = 0.9572$. The other metrics (MAE and $RMSE$) and the cross-validation also showed the good prediction accuracy of the model. On the basis of these results, in addition to the time-prediction model, we empirically demonstrated the first evidence that an error-rate model held well even for crowdsourced user experiments, even though it has been cautioned that crowdworkers are more error-prone in GUI tasks (Findlater et al. 2017; Komarov, Reinecke, and Gajos 2013).

The simulation to alter N_P and N_{repeat} showed that the prediction accuracy of the error-rate model became better when either of these values was larger. This effect was more clearly observed for the error-rate model than the time- and click-point-variability models. In particular for the time model, the prediction accuracy reached close to the upper limit ($R^2 = 1$) even when the N_P and N_{repeat} were not large, such as the $R^2 > 0.95$ exhibited by ten workers shown in Figure 4b. This suggests that the advantage of crowdsourcing in terms of its easy recruitment of numerous workers is not so critical. In comparison, for the error-rate model,

increasing the N_P was still effective for $N_P \geq 160$.

Because the error rate is computed on the basis of occasionally occurring operations (clicking outside the target), researchers need more data to measure the theoretical value. Thus, our result, i.e., that collecting more data would lead to the theoretical value that a model estimates, is intuitive, but it has never been empirically demonstrated until now. Finally, our research hypothesis, "instead of increasing the number of repetitions per task condition, recruiting more workers is another approach to measure the error rates precisely, which will lead to a good prediction accuracy by the error-rate model," was supported. This is a motivating finding for future studies on evaluating novel error-rate models through crowdsourced user experiments.

Note that we compared the sensitivity of time and error-rate models against N_P and N_{repeat}, but our purpose here was not to claim that (e.g.) Fitts' law is a better model than the error-rate model. As described in the Introduction, an MT is measured in every trial and then averaged after completing a session consisting of N_{repeat} trials, but an ER is computed after each session. Due to this difference, surmising that *the error-rate model is inferior* is not appropriate. Although more participants are needed to obtain a good fitness comparable with Fitts' law, which could be a limitation of the error-rate model, it does not necessarily mean that the model is wrong or inaccurate. Collecting numerous participants can avoid reaching such a mistaken conclusion. This point about making a conclusion based on an experiment with small sample size has been made before (Caine 2016; Kaptein and Robertson 2012), and our results again support the importance of a large sample size. Using crowdsourcing for error-rate model evaluation is a straightforward way to enable the recruitment of hundreds of participants with a reasonable time period, cost, and effort by researchers, which enhances the contribution of crowdsourcing to an undeveloped use application.

Limitations and Future Work

Our claims are limited to the task we chose and its design. We emphasized the usefulness of crowdsourced user experiments for error-rate model evaluation, but we only tested a GUI-task model implemented with mice following the Fitts' law paradigm. Within this scope, we limited the task design to horizontal movements where the effect of target height was negligible. We assume that modified models can predict ERs for more realistic targets such as pointing to circular targets (Bi and Zhai 2016; Yamanaka and Usuba 2020), but this needs further investigation in the future.

The model we examined was for selecting static targets, while recently models for more complicated tasks have been proposed, including those for pointing to automatically moving targets (Huang et al. 2019; Lee et al. 2018; Park and Lee 2018), temporally constrained pointing such as rhythm games (Lee and Oulasvirta 2016; Lee et al. 2018), and tracking a moving target (Yamanaka et al. 2020). We assume that the benefit of using crowdsourcing services to recruit numerous participants can be observed in these complicated tasks more clearly than our 1D pointing task. For example, pointing to a circular moving target needs more task parameters, such as the initial target distance A, its size W, movement speed V, and movement angle θ (Hajri et al. 2011; Huang et al. 2019). Because there are more task-condition combinations than 1D-target pointing, it is difficult to ask the participants to perform many repetitions per task condition, while recruiting numerous workers is easy in crowdsourced user studies. Investigating error rates in text input tasks is another important topic in the HCI field (Banovic et al. 2019; Cui et al. 2020) and would be a potential objective for crowdsourced user experiments.

A technical limitation specifically for our GUI-based experiment was that we could not check if workers really followed the given instruction, such as using mice and operating as rapidly and accurately as possible. Similar concerns have been reported before: for touch pointing tasks with smartphones, researchers could not confirm whether workers tapped a target with their thumb as instructed (Yamanaka, Shimono, and Miyashita 2019). Some other crowdsourcing platforms support an option that task requesters can ask workers to shoot a video when they perform a task, e.g., *UIScope* (http://uiscope.com/en). Still, this would create heavier workloads for both the workers and the experimenters. While these issues could not be completely removed at this time, if they were resolved in the future, the contribution to HCI would be significant.

Conclusion

We ran a crowdsourced user experiment to examine the benefits of recruiting numerous participants for evaluating an error-rate prediction model in a target pointing task, which is one of the most fundamental operations in PC usage. By analyzing the data obtained from 384 workers, we found that our model held well with $R^2 > 0.95$. Cross-validation also supported the good prediction accuracy to the unknown task conditions. In addition, when we randomly selected a limited portion of the entire workers from $N_P = 10$ to 320 and used only a limited number of trial repetitions from $N_{\mathrm{repeat}} = 2$ to 10, we found that the time prediction model (Fitts' law) reached $R^2 > 0.95$ even if both of these values were small, while the error-rate model showed quite low fitness in that case. Thus, we empirically demonstrated that using crowdsourcing services for recruiting many participants is more clearly beneficial for evaluating the error-rate prediction model. Our findings should enhance the contribution of crowdsourcing in the HCI field.

References

Bailly, G.; Lecolinet, E.; and Nigay, L. 2016. Visual Menu Techniques. *ACM Comput. Surv.* 49(4). ISSN 0360-0300. doi:10.1145/3002171. URL https://doi.org/10.1145/3002171.

Banovic, N.; Sethapakdi, T.; Hari, Y.; Dey, A. K.; and Mankoff, J. 2019. The Limits of Expert Text Entry Speed on Mobile Keyboards with Autocorrect. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368254. doi:10.1145/3338286.3340126.

Batmaz, A. U.; and Stuerzlinger, W. 2021. The Effect of Pitch in Auditory Error Feedback for Fitts' Tasks in Virtual Reality Training Systems. In *Conference on Virtual Reality and 3D User Interfaces*, VR '21.

Bi, X.; and Zhai, S. 2016. Predicting Finger-Touch Accuracy Based on the Dual Gaussian Distribution Model. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, 313–319. New York, NY, USA: ACM. ISBN 978-1-4503-4189-9. doi:10.1145/2984511.2984546.

Brogmus, G. E. 1991. Effects of Age and Sex on Speed and Accuracy of Hand Movements: And the Refinements they Suggest for Fitts' Law. *Proceedings of the Human Factors Society Annual Meeting* 35(3): 208–212. doi:10.1177/154193129103500311.

Caine, K. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '16, 981–992. New York, NY, USA: Association for Computing Machinery. ISBN 9781450333627. doi:10.1145/2858036.2858498.

Casiez, G.; and Roussel, N. 2011. No More Bricolage!: Methods and Tools to Characterize, Replicate and Compare Pointing Transfer Functions. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, 603–614. New York, NY, USA: ACM. ISBN 978-1-4503-0716-1. doi:10.1145/2047196.2047276.

Cockburn, A.; Lewis, B.; Quinn, P.; and Gutwin, C. 2020. Framing Effects Influence Interface Feature Decisions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–11. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080. doi:10.1145/3313831.3376496.

Crossman, E. R. F. W. 1956. *The speed and accuracy of simple hand movements*. Ph.D. thesis, University of Birmingham.

Cui, W.; Zhu, S.; Zhang, M. R.; Schwartz, H. A.; Wobbrock, J. O.; and Bi, X. 2020. JustCorrect: Intelligent Post Hoc Text Correction Techniques on Smartphones. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, 487–499. New York, NY, USA: Association for Computing Machinery. ISBN 9781450375146. doi:10.1145/3379337.3415857.

Devore, J. L. 2011. *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole, 8th edition. ISBN-13: 978-0-538-73352-6.

Do, S.; Chang, M.; and Lee, B. 2021. A Simulation Model of Intermittently Controlled Point-and-Click Behaviour. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966. doi:10.1145/3411764.3445514.

Findlater, L.; Zhang, J.; Froehlich, J. E.; and Moffatt, K. 2017. Differences in Crowdsourced vs. Lab-based Mobile and Desktop Input Performance Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 6813–6824. New York, NY, USA: ACM. ISBN 978-1-4503-4655-9. doi:10.1145/3025453.3025820.

Fitts, P. M. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47(6): 381–391. doi:10.1037/h0055392.

Gould, S. J. J.; Cox, A. L.; and Brumby, D. P. 2016. Diminished Control in Crowdsourcing: An Investigation of Crowdworker Multitasking Behavior. *ACM Trans. Comput.-Hum. Interact.* 23(3). ISSN 1073–0516. doi:10.1145/2928269.

Grossman, T.; and Balakrishnan, R. 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*, CHI '05, 281–290. New York, NY, USA: Association for Computing Machinery. ISBN 1581139985. doi:10.1145/1054972.1055012.

Hajri, A. A.; Fels, S.; Miller, G.; and Ilich, M. 2011. Moving Target Selection in 2D Graphical User Interfaces. In Campos, P.; Graham, N.; Jorge, J.; Nunes, N.; Palanque, P.; and Winckler, M., eds., *Human-Computer Interaction – INTERACT 2011*, 141–161. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-23771-3.

Hoffmann, E. R. 1997. Movement time of right- and left-handers using their preferred and non-preferred hands. *International Journal of Industrial Ergonomics* 19(1): 49–57. ISSN 0169-8141. doi:10.1016/0169-8141(95)00092-5.

Huang, J.; Tian, F.; Fan, X.; Tu, H.; Zhang, H.; Peng, X.; and Wang, H. 2020. Modeling the Endpoint Uncertainty in Crossing-Based Moving Target Selection. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080. doi:10.1145/3313831.3376336.

Huang, J.; Tian, F.; Fan, X.; Zhang, X. L.; and Zhai, S. 2018. Understanding the Uncertainty in 1D Unidirectional Moving Target Selection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356206. doi:10.1145/3173574.3173811.

Huang, J.; Tian, F.; Li, N.; and Fan, X. 2019. Modeling the Uncertainty in 2D Moving Target Selection. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, 1031–1043. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368162. doi:10.1145/3332165.3347880.

Jokinen, J. P. P.; Sarcar, S.; Oulasvirta, A.; Silpasuwanchai, C.; Wang, Z.; and Ren, X. 2017. *Modelling Learning of New Keyboard Layouts*, 4203–4215. New York, NY, USA: Association for Computing Machinery. ISBN 9781450346559. URL https://doi.org/10.1145/3025453.3025580.

Kaptein, M.; and Robertson, J. 2012. Rethinking Statistical Analysis Methods for CHI. In *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems, CHI '12, 1105–1114. New York, NY, USA: Association for Computing Machinery. ISBN 9781450310154. doi:10.1145/2207676.2208557.

Komarov, S.; Reinecke, K.; and Gajos, K. Z. 2013. Crowdsourcing Performance Evaluations of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 207–216. New York, NY, USA: ACM. ISBN 978-1-4503-1899-0. doi:10.1145/2470654.2470684.

Lee, B.; Kim, S.; Oulasvirta, A.; Lee, J.-I.; and Park, E. 2018. Moving Target Selection: A Cue Integration Model. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 230:1–230:12. New York, NY, USA: ACM. ISBN 978-1-4503-5620-6. doi:10.1145/3173574.3173804.

Lee, B.; and Oulasvirta, A. 2016. Modelling Error Rates in Temporal Pointing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 1857–1868. New York, NY, USA: ACM. ISBN 978-1-4503-3362-7. doi:10.1145/2858036.2858143.

MacKenzie, I. S. 1992. Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction* 7(1): 91–139. doi:10.1207/s15327051hci0701_3.

MacKenzie, I. S. 2002. Bibliography of Fitts' Law Research. Retrieved August 24, 2021 from http://www.yorku.ca/mack/RN-Fitts_bib.htm.

MacKenzie, I. S.; and Isokoski, P. 2008. Fitts' Throughput and the Speed-Accuracy Tradeoff. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, 1633–1636. New York, NY, USA: ACM. ISBN 9781605580111. doi:10.1145/1357054.1357308.

MacKenzie, I. S.; Kauppinen, T.; and Silfverberg, M. 2001. Accuracy Measures for Evaluating Computer Pointing Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, 9–16. New York, NY, USA: ACM. ISBN 1-58113-327-8. doi:10.1145/365024.365028.

Matejka, J.; Glueck, M.; Grossman, T.; and Fitzmaurice, G. 2016. The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 5421–5432. New York, NY, USA: Association for Computing Machinery. ISBN 9781450333627. doi:10.1145/2858036.2858063.

Meyer, D. E.; Abrams, R. A.; Kornblum, S.; Wright, C. E.; and Smith, J. E. K. 1988. Optimality in human motor performance: ideal control of rapid aimed movements. *Psychological Review* 95(3): 340–370. doi:10.1037/0033-295X.95.3.340.

Park, E.; and Lee, B. 2018. Predicting Error Rates in Pointing Regardless of Target Motion. URL https://arxiv.org/abs/1806.02973.

Plamondon, R.; and Alimi, A. M. 1997. Speed/accuracy trade-offs in target-directed movements. *Behavioral and Brain Sciences* 20(2): 279–303. doi:10.1017/S0140525X97001441.

Schwab, M.; Hao, S.; Vitek, O.; Tompkin, J.; Huang, J.; and Borkin, M. A. 2019. Evaluating Pan and Zoom Timelines and Sliders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359702. doi:10.1145/3290605.3300786.

Soukoreff, R. W.; and MacKenzie, I. S. 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. doi:10.1016/j.ijhcs.2004.09.001.

Wobbrock, J. O.; Cutrell, E.; Harada, S.; and MacKenzie, I. S. 2008. An Error Model for Pointing Based on Fitts' Law. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, 1613–1622. New York, NY, USA: ACM. ISBN 978-1-60558-011-1. doi:10.1145/1357054.1357306.

Wobbrock, J. O.; Findlater, L.; Gergle, D.; and Higgins, J. J. 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. New York, NY, USA: ACM. ISBN 978-1-4503-0228-9. doi:10.1145/1978942.1978963.

Yamanaka, S. 2021. Comparing Performance Models for Bivariate Pointing through a Crowdsourced Experiment. In *Human-Computer Interaction – INTERACT 2021*, 76–92. Gewerbestr, Switzerland: Springer International Publishing. ISBN 978-3-030-85616-8. doi:10.1007/978-3-030-85616-8_6.

Yamanaka, S.; Shimono, H.; and Miyashita, H. 2019. Towards More Practical Spacing for Smartphone Touch GUI Objects Accompanied by Distractors. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*, ISS '19, 157–169. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368919. doi:10.1145/3343055.3359698.

Yamanaka, S.; and Usuba, H. 2020. Rethinking the Dual Gaussian Distribution Model for Predicting Touch Accuracy in On-Screen-Start Pointing Tasks. *Proc. ACM Hum.-Comput. Interact.* 4(ISS). doi:10.1145/3427333.

Yamanaka, S.; Usuba, H.; Takahashi, H.; and Miyashita, H. 2020. Servo-Gaussian Model to Predict Success Rates in Manual Tracking: Path Steering and Pursuit of 1D Moving Target. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, 844–857. New York, NY, USA: Association for Computing Machinery. ISBN 9781450375146. doi:10.1145/3379337.3415896.

Yu, D.; Liang, H.-N.; Lu, X.; Fan, K.; and Ens, B. 2019. Modeling Endpoint Distribution of Pointing Selection Tasks in Virtual Reality Environments. *ACM Trans. Graph.* 38(6). ISSN 0730-0301. doi:10.1145/3355089.3356544.

Zhai, S.; Kong, J.; and Ren, X. 2004. Speed-accuracy tradeoff in Fitts' law tasks: on the equivalency of actual and nominal pointing precision. *International Journal of Human-Computer Studies* 61(6): 823–856. doi:10.1016/j.ijhcs.2004.09.007.

Zhao, J.; Soukoreff, R. W.; Ren, X.; and Balakrishnan, R. 2014. A model of scrolling on touch-sensitive displays. *International Journal of Human-Computer Studies* 72(12): 805 – 821. ISSN 1071-5819. doi:10.1016/j.ijhcs.2014.07.003.