Comparing Performance Models for Bivariate Pointing Through a Crowdsourced Experiment

Shota Yamanaka

Comparing Performance Models for Bivariate Pointing through a Crowdsourced Experiment

Shota Yamanaka

Yahoo Japan Corporation (syamanak@yahoo-corp.jp)

Abstract. Evaluation of a novel user-performance model's fitness requires comparison with baseline models, yet it is often time consuming and involves much effort by researchers to collect data from many participants. Crowdsourcing has recently been used for evaluating novel interaction techniques, but its potential for model comparison studies has not been investigated in detail. In this study, we evaluated four existing Fitts' law models for rectangular targets, as though one of them was a proposed novel model. We recruited 210 crowd workers, who performed 94,080 clicks in total, and confirmed that the result for the best-fit model was consistent with previous studies. We also analyzed whether this conclusion would change depending on the sample size, but even when we randomly sampled data from five workers for 10,000 iterations, the best-fit model changed only once (0.01%). We have thus demonstrated a case in which crowdsourcing is beneficial for comparing performance models.

Keywords: Performance modeling · Fitts' law · Crowdsourcing.

1 Introduction

It has recently become common for researchers to employ workers through crowdsourcing services for user experiments on graphical user interfaces (GUIs) [8, 12, 21, 25, 34]. However, those works focused mainly on designing better GUIs and evaluating novel interaction techniques in comparison with baseline methods. In this paper, we explore the potential utility of crowdsourced experiments to evaluate user performance models. Deriving a novel model to predict operation times on GUIs is a common topic in the human-computer interaction (HCI) field, but model evaluation is typically conducted in lab-based experiments with 10 or 20 university students, i.e., a limited subset of all computer users. If we instead used crowdsourcing for model comparison, it would save time for researchers and improve the evaluation validity because of the large number and diversity of the participants.

It is unclear, however, that we can use crowdsourcing as an alternative to lab-based experiments. For example, crowd workers use different mice, displays, operating systems (OSs), and so on. Also, it is known that there are many digital and non-digital distractors that can break workers' focus [17]. Even so, can crowdsourced experiments give the same conclusions on model evaluation as lab-based experiments? To investigate this, we replicated a model-comparison

experiment that was previously conducted through lab-based experiments; i.e., the answer on the best model is already known. This enabled us to examine whether we could reach the same conclusion obtained in the lab-based experiments.

Specifically, we examined a bivariate (rectangular) target pointing task, which is modeled by modified versions of Fitts' law [13]. Because Fitts' law tasks have a well-structured methodology, are easy to conduct in desktop environments, and take a short time (typically less than 10 min), they are suitable for crowdsourced user experiments. Also, several formulations have been proposed for bivariate pointing tasks, but Accot and Zhai's weighted Euclidean model [2] is already considered the best-fit model; thus, we could determine whether a crowdsourced experiment would lead to the same conclusion on the best model. In other words, if Accot and Zhai's model were "our proposed novel model" and we identified it as the best in a crowdsourced experiment, we could conclude that our finding on the best model was consistent with the lab-based finding.

Our contributions are as follows.

- We conducted a crowdsourced mouse-pointing experiment with rectangular targets. In total, we recorded 94,080 clicks performed by 210 crowd workers. In line with previous studies, we confirmed that Accot and Zhai's model showed the best fit (adjusted $R^2 = 0.9631$).
- We simulated how the number of participants, N_P, affected the conclusion on the best-fit model. By randomly sampling worker data for N_P values from 5 to 100 (interval: 5) and testing the model fitness over 100 iterations, we found that the best model never changed. Because the model fitness had larger variability when the N_P was smaller, we also performed this simulation with $N_P = 5$ over 10,000 iterations. Even in that case, the best-fit model changed only once (i.e., with a 0.01% chance), which showed the robustness of crowdsourced model comparison even for a small sample size, at least in one case (bivariate pointing).

2 Related Work

2.1 Fitts' Law and Modified Versions for Bivariate Pointing

Fitts' law expresses the notion that the movement time MT to point to a target is related to the index of difficulty in bits, ID, as follows [13]:

$$MT = a + b \cdot ID, \tag{1}$$

where a and b are empirical regression constants. The Shannon formulation of the ID [22] is widely used in HCI:

$$ID = \log_2 \left(\frac{A}{W} + 1 \right), \tag{2}$$

where the distance between the target centers is A, and the target size (or width) is W. As shown in Figure 1a, a traditional Fitts' law task has two ribbon-shaped

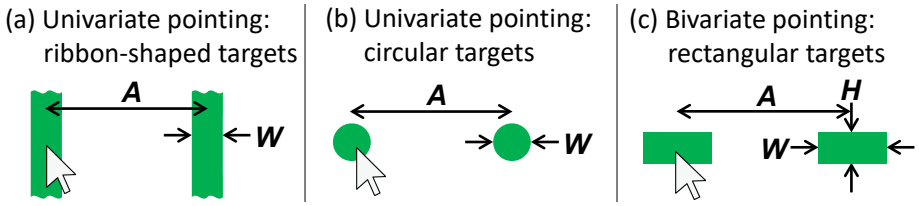

Fig. 1. Pointing tasks with different target shapes.

targets; thus, participants do not need to pay attention to the cursor's y-axis movements. It is also common to use circular targets, which constrain the y-axis movements, as shown in Figure 1b. In both (a) and (b), the target shape is defined by its width W alone; i.e., the target is univariate. In contrast, more realistic targets such as buttons and icons have another dimension, the height H; i.e., they are bivariate targets, as shown in Figure 1c.

Crossman proposed the first model to predict the MT for such rectangular targets, which used another empirical regression constant c [9]:

$$MT = a + b \cdot \log_2\left(\frac{A}{W} + 1\right) + c \cdot \log_2\left(\frac{A}{H} + 1\right). \tag{3}$$

To make the regression expression clearer, we modify this model as follows:

$$MT = a + b \cdot \left[\log_2\left(\frac{A}{W} + 1\right) + c' \cdot \log_2\left(\frac{A}{H} + 1\right)\right], \tag{4}$$

where $c' = c/b$ (b cannot be zero). Crossman's original formulation did not include the "+1" factors. For fair comparison with other models, however, we consistently include these "+1" factors, as in Accot and Zhai's work [2]. This decision does not affect our conclusion because it has little effect on model fitness [16, 27].

MacKenzie and Buxton [23] and Hoffmann and Sheikh [19] independently proposed the same model using the smaller of W and H:

$$MT = a + b \cdot \log_2\left(\frac{A}{\min(W, H)} + 1\right). \tag{5}$$

This model indicates that the time is solely affected by the more difficult dimension. Accot and Zhai proposed another successful model for bivariate pointing, called the weighted Euclidean model:

$$MT = a + b \cdot \log_2\left(\sqrt{\left(\frac{A}{W}\right)^2 + c \cdot \left(\frac{A}{H}\right)^2} + 1\right), \tag{6}$$

where c is a weight for the target height with respect to the width. Hoffmann et al. identified this model as the best for bivariate pointing with a physical stylus [18], while Accot and Zhai used a mouse.

There are other variations for bivariate pointing, such as integrating the cursor's movement angle factor θ [4, 23, 36, 38]. To limit our focus in this study, and to limit the task time for crowd workers, our experiment used only single-axis movements. This choice is consistent with previous studies [2, 9, 18, 19].

2.2 Crowdsourced Studies on GUI Tasks and Model Evaluations

There have been reports on the consistency between lab-based and crowdsourced experiments involving GUI operations. For menu selection and target pointing tasks in desktop environments, Komarov et al. found that crowdsourced and lab-based experiments gave the same findings on user performance, such as the finding that novel techniques were better than baseline operations [21]. Yamanaka et al. tested the effects of target margins on touch-pointing tasks, and they reported that the same effects were consistently found in crowdsourced and lab-based experiments. For example, in both kinds of experiments, wider margins decreased the MT but increased the error rate [34]. In contrast, by using more powerful statistical analysis methods and recruiting many more participants for lab-based experiments, Findlater et al. showed that crowd workers had significantly shorter average task completion times and higher average error rates in both mouse- and touch-pointing tasks [12]. Thus, they cautioned against assuming that crowdsourced data from GUI performance experiments directly reflects lab-based data.

As for Fitts' law fitness analyses, Findlater et al. reported that crowd workers had average values of $r = 0.926$ with mice and $r = 0.898$ with touchscreens [12]. Schwab et al. conducted a crowdsourced scrolling task in desktop and mobile environments [28]. The results showed that Fitts' law fit the operation times with $R^2 = 0.983$ and 0.972 for the desktop and mobile cases, respectively (note that scrolling operations follow Fitts' law well [39]). Overall, these reports suggest that Fitts' law is valid for crowdsourced data, regardless of the operation style.

To our knowledge, the only literature on using crowdsourcing to determine a best-fit model was the work by Goldberg et al. [15][1]. They implemented Fitts' law tasks in an applet on their website and let visitors to the site perform the tasks. More than 5,000 visitors performed 78,410 clicks in total. Their focus was on whether the best-fit model would change depending on the A/W ratio. For example, when A/W was less than 5, Meyer et al.'s model ($ID = \sqrt{A/W}$) [26] was significantly better, while for harder tasks, the Shannon formulation (Equation 2) was better.

There are several differences in focus between our work and Goldberg et al.'s work. First, they were interested in how model fitness differences depend on the task difficulty. This is important for understanding user behaviors in pointing tasks [14, 26], but typically, model fitness is evaluated in terms of the regression expression for all A/W data points. Also, they mainly compared Meyer et al.'s

[1] We found this previous work as part of a Ph.D. thesis by one of these authors (Faridani) [11]. He defined this Fitts' law study as a crowdsourced task, and thus we introduce it here.

square-root-based model with the Shannon (logarithmic) model, which can be mathematically approximated [27]. Because their comparison cannot clearly reveal the effectiveness of crowdsourced model-comparison experiments, it is not applicable to our purpose in this study. Moreover, their participants were unpaid volunteers, and thus they called their study an "uncontrolled user study." In comparison, we paid our workers and thus assumed they were motivated to follow our experimental controls (i.e., instructions). Therefore, our results are more relevant for researchers who use crowdsourcing services such as Amazon MTurk.

In summary, no previous studies are directly related to our research question of how useful crowdsourced user experiments are for comparing novel performance models with baselines. If we can demonstrate the potential of crowdsourced model comparison, at least for one example task (bivariate pointing), it will enable future researchers to investigate novel performance models with less recruitment effort, more diversity of participants, and less time-consuming data collection; this is our motivation for this work.

3 Experiment

We conducted a pointing experiment with rectangular targets by using *Yahoo! Crowdsourcing*[2]. The experimental system was developed with the `Hot Soup Processor` programming language (version 3.5). The crowd workers were asked to download and run an executable file from a URL on the recruitment page.

3.1 Task, Design, and Procedure

The task was to click a red target that had width W and height H. The study was a 4×7 within-subjects design: four W values (30, 40, 60, and 90 pixels) and seven H values (10, 20, 30, 40, 60, 100, and 200 pixels). The target distance was fixed to 640 pixels to limit the number of task condition combinations. Using only one A value is consistent with previous studies on bivariate pointing [6, 9, 19]. A *session* consisted of 21 cyclic clicks back and forth between the left and right targets with a fixed $W \times H$ condition. Each participant completed 28 ($= 4_W \times 7_H$) sessions.

The first target was on the left side. If the participant clicked the target, the red target and white nontarget rectangles switched colors, as illustrated in Figure 2a. If the participant missed the target, it flashed yellow, and the participant had to keep trying until he or she successfully clicked 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. After completing 21 successful clicks, the participant saw the results of the session and a message to take a break, as shown in Figure 2b.

After finishing 28 sessions, the participants completed a questionnaire on their age (numeric), gender (free-form to allow nonbinary or arbitrary answers),

[2] https://crowdsourcing.yahoo.co.jp

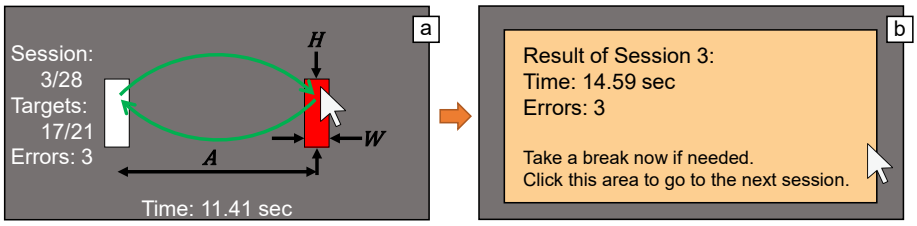

Fig. 2. (a) In the task, participants clicked alternately on each target when it was red. (b) At the end of a session, the results and a message to take a break were shown.

handedness (left or right), Windows version (free-form), input device (free-form), and history of PC use (numeric in years). The questionnaire also included a free-form response for comments on the task and their impressions.

To measure the central tendency of each participant's performance, it is recommended to require 15 to 25 clicks under each condition [30]. Thus, we treated the first five clicks as practice and the remaining 16 clicks (8 clicks on each target) as data. The order of the 28 $W \times H$ conditions was randomized. In total, we recorded $4_W \times 7_H \times 16_{\text{clicks}} \times 210_{\text{workers}} = 94{,}080$ data points.

3.2 Participants and Recruitment

We recruited workers who used Windows Vista or a later version to run our system. We used the "white list" option in the crowdsourcing platform to screen newly created accounts and prevent multiple entries. This option enabled us to offer the task only to workers who were considered reliable from their previous task history; however, criteria such as the approval rate were not available on the platform.

In the recruitment page, we asked the workers to use a mouse if possible. We made this request because, in our data analysis, we randomly selected a number of participants (e.g., 10) to examine the model fitness. If different devices were used (e.g., six mice, two touchpads, and two trackballs), we might have wondered if a poor fit was due to the device differences. Nevertheless, to avoid a possible false report that all the workers used mice, we did not limit them to using mice. Instead, we specified that any device was acceptable and then removed the non-mouse users from the analysis.

Once a worker accepted the task, he or she was asked to read the online instructions, which stated that the worker should perform the task as rapidly and accurately as possible. After a worker finished all 28 sessions and completed the questionnaire, the log data was exported to a csv file. The worker uploaded the file to a server and then received a payment of JPY 100 ($\approx$ USD 0.96). The main pointing task typically took 8 or 9 minutes to complete, and thus the effective hourly payment was approximately JPY 700 ($\approx$ USD 6.7).

In total, 225 workers completed the task, including 210 mouse users. The mouse users' demographics were as follows. Age: 20 to 64 years, with M =

43.5 and $SD = 8.70$. Gender: 160 male, 47 female, and 3 chose not to answer. Handedness: 18 were left-handed and 192 were right-handed. Windows version: 26 used Win7, 4 used Win8, 4 used Win8.1, and 176 used Win10. PC usage history: 3 to 47 years, with $M = 21.0$ and $SD = 6.86$.

4 Results

4.1 Outlier Data Screening

Following previous studies [12, 24], we removed spatial trial-level outliers if (1) the distance of the first click position was shorter than $A/2$ or (2) the click position of the x-coordinate was more than $2W$ away from the target center. We also applied the latter criterion to the y-coordinate: trials in which the click position was more than $2H$ away from the target center were removed.

To detect trial-level temporal outliers, we used a robust means of outlier detection called the inter-quartile range (IQR) method [10]. The IQR is defined as the difference between the third and first quartiles of the MT. Trials in which the MT was more than $3IQR$ higher than the third quartile or more than $3IQR$ lower than the first quartile were removed. This calculation was run for each session.

For participant-level outliers, we calculated the mean MT across all 28 conditions ($4_W \times 7_H$) for each participant. Then, using each participant's mean MT, we again applied the IQR method. Note that the trial- and participant-level outliers were independently detected and removed.

As a result, among the 94,080 trials, we found 1,043 trial-level outliers (1.11%). We also found one participant-level outlier worker. While the other participants' mean MT was 838 ms, this worker's mean MT was 1,487 ms, and nine of the worker's trials had $MT >$ 3,000 ms. Accordingly, all 448 ($= 16_{\text{clicks}} \times 4_W \times 7_H$) data points for this worker were removed. He or she also had trial-level outliers (i.e., there were overlaps); as a result, 1,487 data points were removed in total (1.58%).

4.2 Analyses of Dependent Variables

After the outliers were removed, 92,593 data points (98.4%) were analyzed. The dependent variables were the error-free MT and the error rate ER. Hereafter, any MT value represents error-free data.

Movement Time The Shapiro-Wilk test ($\alpha = 0.05$) showed that among the $4_W \times 7_H \times 209_{\text{workers}} = 5852$ conditions, 4983 MT data points passed the normality test (85.2%). Hence, to meet the normality assumption, we log-transformed the data before applying repeated-measures ANOVA. We used 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 ($\alpha = 0.05$).

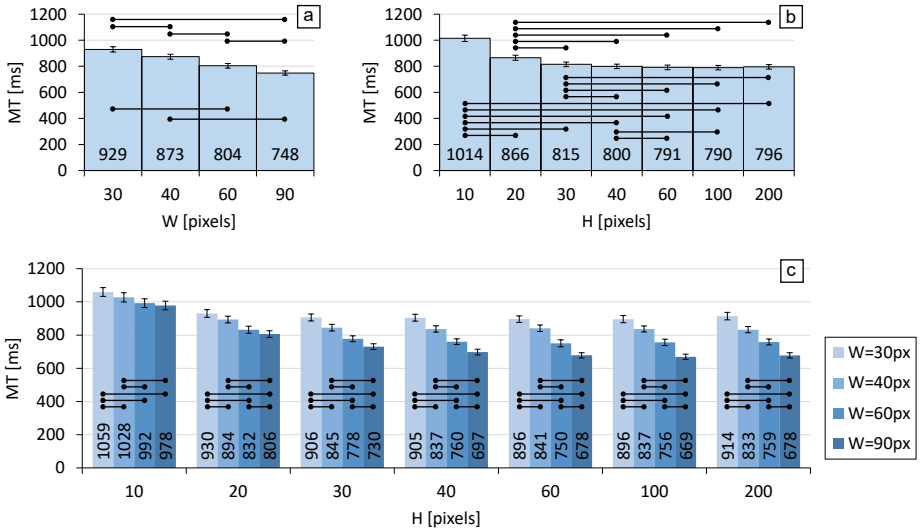

Fig. 3. (a, b) Main effects and (c) interaction for the MT. The error bars show 95% CIs. The horizontal bars show significant differences ($p < 0.05$ at least).

We found significant main effects of W ($F_{2.540,528.243} = 1876.778$, $p < 0.001$, $\eta_p^2 = 0.900$) and H ($F_{3.247,675.285} = 851.453$, $p < 0.001$, $\eta_p^2 = 0.704$) on MT. A significant interaction was found for $W \times H$ ($F_{13.603,2829.404} = 52.468$, $p < 0.001$, $\eta_p^2 = 0.201$). As the W increased, the MT decreased, and all pairwise tests (six combinations) showed significant differences (Figure 3a). In contrast, the effect of H on MT plateaued gradually, and the pairwise tests for $H \geq 60$ showed no significant differences (Figure 3b).

As for the interaction effect, for the largest H (200 pixels), the effect of W on MT was more clearly observed, as seen in Figure 3c; this means that the effect of W was dominant. As the H decreased, however, the MT differences among the four W values were reduced, because H was dominant. For example, the largest difference for $H = 10$ pixels was $1059 - 978 = 81$ ms (7.6%), while that for $H = 200$ pixels was $914 - 678 = 236$ ms (26%). This result demonstrates that we should integrate the interaction effect of $W \times H$ to predict the MT accurately.

Error Rate Error-rate data are typically nonparametric; thus, we used a nonparametric ANOVA with the *Aligned Rank Transform* [31] and Tukey's p-value adjustment method for pairwise tests. We found significant main effects of W ($F_{3,624} = 15.146$, $p < 0.001$, $\eta_p^2 = 0.068$) and H ($F_{6,1248} = 49.095$, $p < 0.001$, $\eta_p^2 = 0.191$) on ER. A significant interaction was found for $W \times H$ ($F_{18,3744} = 3.670$, $p < 0.001$, $\eta_p^2 = 0.017$). As the W increased, the MT gradually decreased (Figure 4a), while for the H, the ER for 10 pixels was remarkably high, as seen in (b). For the interaction effect (c), when the H was small (10 or 20 pixels), the pairwise tests for W showed no effects, which indicates the dominance of H.

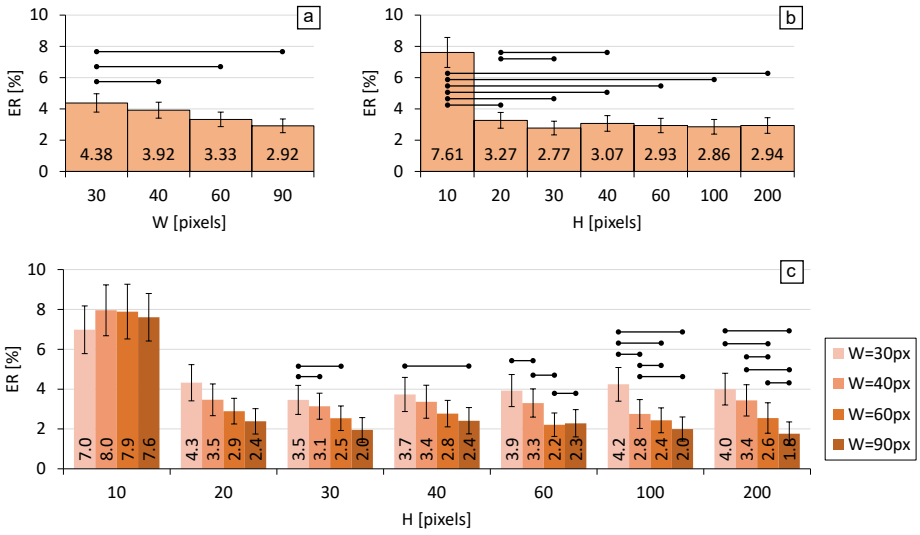

Fig. 4. (a, b) Main effects and (c) interaction for the ER. The error bars show 95% CIs. The horizontal bars show significant differences ($p < 0.05$ at least).

Then, as the H increased, it lost the effect to impose errors; thus, the W had the dominant effect on ER.

4.3 Canonical analysis

In addition to the ANOVAs, we ran a canonical analysis to examine how $\{W, H\}$ affected the two dependent variables of $\{MT, ER\}$ concurrently. We used the `CCA` function provided by the `sklearn.cross_decomposition` library in Python. The Pearson's r values for the first and second dimensions were 0.8974 and 0.4238, respectively. For the independent variables, the canonical loadings were $[0.9962, 0.0873]$, $[-0.0873, 0.9962]$. Thus, W had a stronger effect on the dependent variables than H; this is consistent with previous studies [2, 18]. For the dependent variables, the canonical loadings were $[-0.9385, 0.4091]$, $[-0.5869, -0.8096]$. Thus, MT was affected by the independent variables more sensitively than ER. As shown in Figure 4a–b, while the $H = 10$ pixels condition is an exception, W and H did not largely change the ER.

4.4 Model Fitness

Figure 5 summarizes the results of model fitness for the four candidate models. Because the numbers of free parameters in the models (a, b, and c) are different, it was necessary to use the *adjusted* R^2 rather than R^2. In addition, to compare the model fitness more statistically, the figure shows the Akaike information criterion (AIC) values [3]. The AIC enabled us to determine comparatively

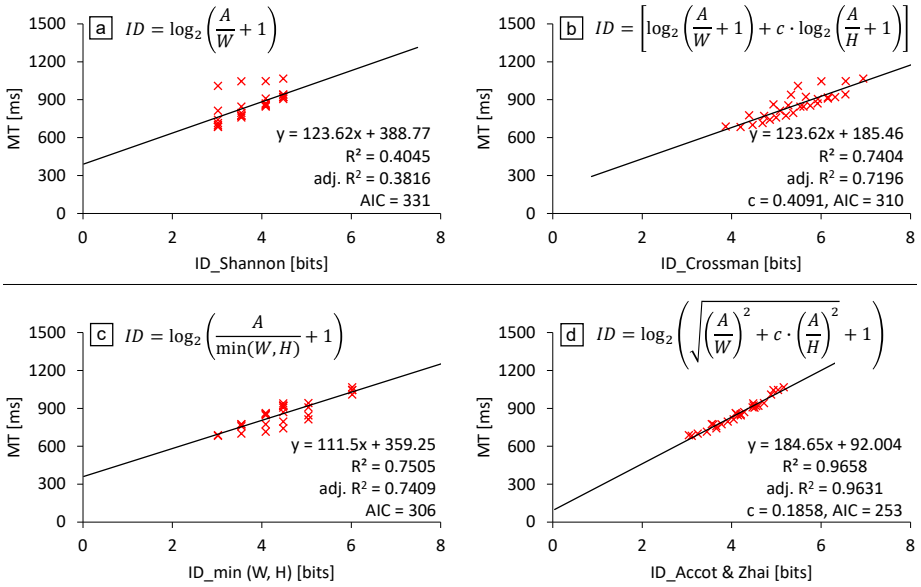

Fig. 5. Regression results for the four candidate models: (a) Shannon from Equation 2, (b) Crossman from Equation 4, (c) "$\min(W, H)$" from Equation 5, and (d) weighted Euclidean from Equation 6.

better models in terms of the number of parameters, via the following brief rule of thumb: (a) a model with a lower AIC value is better, and the one with the minimum AIC (AIC_{minimum}) is thus the best; (b) a model with $AIC \leq (AIC_{\text{minimum}} + 2)$ is comparable with the better models; and (c) a model with $AIC \geq (AIC_{\text{minimum}} + 10)$ can be safely rejected [7]. To simplify the discussion in this paper, we consider an AIC difference greater than 10 to be significant.

As shown in Figure 5a, the baseline model (Shannon) could not capture the fact that the MT depended on the H, because it considers only the W. The data points' vertical spread shows that the H significantly affected the MT. The modified Crossman and "$\min(W, H)$" models (Figure 5b and c, respectively) showed improved model fitness. The points spread farther horizontally, and more points thus lay on the regression lines as compared with the Shannon model. The adjusted R^2 values increased from 0.38 to at least 0.7, and the AIC values significantly decreased from 331 to 310 or lower.

Among the four candidates, Accot and Zhai's model showed the best fit for both the adjusted R^2 and AIC values (0.9631 and 253, respectively). All the points lay close to the regression line, and we could thus visually confirm that the MT can be predicted most accurately with this model.

4.5 Answers to the Free-form Questionnaire

Among the 210 workers, 15 mentioned the effects of target size on task difficulty, e.g., "Small bars were difficult to click." Also, two workers explicitly mentioned the target height: "Horizontally long bars were more difficult to click than vertically long ones," which clearly indicates that the effect of H on task performance was dominant rather than W. Five workers stated that "It was easier to click horizontally long buttons" and one of them stated that "The frequency of clicking outside targets was lower for horizontally long bars," which indicates that a larger W had a positive effect on lowering the error rate. This result partially supports the result that the weight for W was heavier than H in model fitting (i.e., in Figure 5d, the weight for W was 1 vs. 0.1858 for H).

4.6 Discussion on Model Fitness

Through this experiment, we confirmed that our crowdsourced data gave the same conclusion obtained in lab-based experiments. However, this result might have depended on the large number of participants: for example, even if some participants ignored the instructions [12] or did multiple tasks [17], we can compress such noisy data after recruiting over 200 workers. This easy recruitment is an advantage of crowdsourced user experiments, but it means that researchers might have to pay a high cost to obtain less noisy data. If we could reach the same conclusion with a much lower number of workers, such as 10% of the total here, it would also demonstrate the utility of crowdsourcing services. However, it is currently unclear how the sample size N_P affects the conclusion on model fitness. To assess this issue, we ran the simulation study described in the next section.

5 Simulation of Sample Size Effect on Model Fitness

Through this simulation, we analyzed how the number of participants, N_P, affects the model fitness and the conclusion on the best-fit model. We randomly selected N_P participants' data from the 210 crowd workers and computed the model fitness in terms of the adjusted R^2 and AIC. To handle the randomness, we repeated the simulation over 100 iterations for a single N_P value. Then, we computed the average and SD of the adjusted R^2 and AIC for the 100 iterations[3].

Figure 6 shows the results of this simulation, in which we varied the N_P from 5 to 100 with an interval of 5 (i.e., $20_{N_P} \times 100_{\text{iterations}} = 2000$ simulation trials). Regardless of the N_P, Accot and Zhai's model (green lines) was the best in terms of both the adjusted R^2 and the AIC. In addition, we could visually confirm that the mean adjusted R^2 values became stable after the N_P reached

[3] The simulation included data from the outlier worker detected in the analysis of the main experiment, because that worker's status as an outlier depends on the other sampled workers' results.

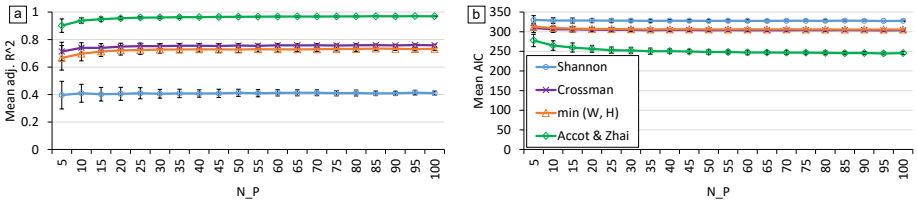

Fig. 6. Model fitness in terms of the (a) adjusted R^2 and (b) AIC, depending on the sample size. Each point shows the $Mean \pm 1SD$ obtained through 100 iterations.[4]

approximately 20 or 25 for all the model candidates. This was also true for the AIC result. Throughout the 2000 simulation trials, Accot and Zhai's model showed the lowest AIC values, and the difference from the second-best model was always greater than 10; i.e., Accot and Zhai's model was consistently and significantly the best.

According to Figure 6, as the N_P decreased, the variabilities in the adjusted R^2 and AIC increased (larger error bars). Thus, we assumed the possibility that one of the other three candidates could become the best-fit model. To examine this assumption, we ran the simulation again with $N_P = 5$ over 10,000 iterations. The results showed only one trial in which Accot and Zhai's model had a lower adjusted R^2 value than the $\min(W, H)$ model: 0.559 vs. 0.592, respectively. For the AIC, however, there was no significant difference: 308 vs. 305. In contrast, in 9,998 simulation trials, Accot and Zhai's model was significantly the best model according to the AIC.

In conclusion, note that it would have been possible to observe the "opposite" conclusion from lab-based experiments, in which Accot and Zhai's model is not the best, but the probability of that situation was only 0.01%. Also, as the N_P increases, this probability should approach zero according to our simulation, which showed that the variabilities in model fitness became quite small.

Ideally, we would try all combinations of selecting $N_P = 5$ participants' data among the 210 crowd workers. However, our simulation of only 10,000 iterations took approximately 28 min; testing the ${}_{210}\mathrm{C}_5 = 3{,}244{,}032{,}792$ combinations for this case would take 17 years, which is not feasible. Also, the simulation took longer times with larger N_P values for various reasons, such as random selection, averaging of more MT values, detection of outlier workers, and nonlinear regressions; thus, we tested only the remarkable case of the greatest variability in model fitness.

[4] In addition to the mean and SD, we computed the [min, max] values and 95% CIs [lower, upper] of the adjusted R^2 and AIC. The 95% CI was used for estimating the true value, but our goal here is to discuss how the sample size affects the mean and the variability of model fitness; thus, we show the $Mean \pm 1SD$ in this figure.

6 General Discussion

6.1 Benefits of Using Crowdsourcing for Model Comparison

In this study, we explored the potential of crowdsourcing for GUI operation model evaluation studies in desktop environments. As an example of a fundamental and well-structured experiment, a Fitts' law task with bivariate targets was used. The results obtained from 210 crowd workers showed that the best fit was achieved by the weighted Euclidean model proposed by Accot and Zhai (Equation 6): adjusted $R^2 = 0.9631$ and $AIC = 253$. This conclusion on the best model was consistent with previous studies [2, 18].

Although comparison of GUI operation models through crowdsourcing is not common in HCI research, we have demonstrated its effectiveness, at least for one example (bivariate pointing). This is a motivating result for future studies on evaluating novel user-performance models. Also, according to a follow-up simulation, our conclusion on the best-fit model would not have changed in most cases, even if we had conducted this crowdsourced experiment with only five workers. An experiment that size would cost only JPY 500 ($\approx$ USD 4.8), thus enabling easy model fitness comparison at low cost. Furthermore, as the sample size increased, we observed a more robust, stable model fitness (i.e., less variability in the adjusted R^2 and AIC). Hence, if researchers can pay more to recruit more workers (e.g., $N_P = 100$) to improve the reliability of the data, such large model fitness studies can easily be performed through crowdsourcing, while lab-based experiments of that size are comparatively difficult.

6.2 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 model comparison, but we only tested GUI operation models with mice. Even within the scope of Fitts' law tasks, we purposely limited the task design to horizontal movements and a fixed target distance so that the parameters W and H could be reasonably varied. We will need further studies on the applicability to other kinds of models such as cognitive ones and other input devices such as touchscreens.

Another possible limitation of our data analysis is that we used a single criterion for outlier detection, particularly for spatial outliers. While our criterion was based on $2W$ and $2H$ (see Section 4.1), some previous studies have used different criteria, e.g., $8W$ [37]. Thus, we ran a pseudo-ablation study: the click position of the x-coordinate was more than NW away from the target center (and also for the y-axis), where N ranged from 1 to 10 (the severest outlier criterion to the most relaxed one, respectively). As a result, among the 94,080 trials, we found 1,142 trial-level outliers (1.214%) when $N = 1$. This changed to 1,031 outliers (1.096%) when $N = 10$. This 0.1-point difference did not affect our conclusion. For example, Accot and Zhai's model showed adjusted $R^2 = 0.9729$ and 0.9728 for $N = 1$ and 10, respectively. This additional analysis demonstrates that our outlier detection criteria do not change our main claim.

A crowdsourcing-specific limitation for GUI tasks is that we cannot check if workers really follow the given instruction. For example, a previous study has mentioned that an experimenter could not check whether workers tapped a target with their thumb as instructed (e.g., some workers might have used their index finger when tapping a small target) [34]. Taking this into consideration, we cannot recommend measuring Fitts' law fitness and computing the throughput for comparing the performance of various devices such as mouse vs. trackball [30] because some workers might not use the specified device and the data reliability is thus questionable.

As a more important argument, we tested four model candidates and compared their model fitness as though our proposed model was Accot and Zhai's weighted Euclidean model. Because the results of lab-based experiments are already known (i.e., the finding that Accot and Zhai's model is the best), this study design enabled us to examine whether the crowdsourced user experiment gave the same conclusion. However, if future researchers conduct crowdsourced user experiments to evaluate their novel models with respect to baselines, it is unknown whether the conclusion on the best-fit model will be the same as in lab-based experiments. In this case, there will be no choice but to believe the best model as-is.

Still, such results obtained by crowdsourcing would motivate further lab-based experiments if more controlled conditions (e.g., the same device settings and no interruptions) and reliable participants who follow instructions are needed. Therefore, crowdsourced and lab-based experiments have different characteristics, and our purpose in this paper is not to state a binary claim on which choice is better. Rather, we seek to open up a new possibility of using crowdsourcing as a tool for HCI studies, particularly for human motor performance modeling.

In this study, we focused on a mouse pointing task. Other potential examples to use crowdsourcing for model evaluation include the steering law [1] and its refined versions for other path conditions [29, 33, 32], and Fitts' law for finger touching [5] and its refined versions [20, 35]. Findlater et al. found that Fitts' law held well for touchscreens [12], and thus it is promising to use crowdsourcing for evaluating new models on touchscreen interactions. To demonstrate this generalizability to other tasks, more experiments under different conditions are required.

7 Conclusion

We conducted a crowdsourced user experiment to compare model fitness on a bivariate Fitts' law task. By analyzing the data obtained from 210 crowd workers, we found that the conclusion on the best model was consistent with previous studies. In addition, even when we randomly selected a limited number of workers from 5 to 100, we consistently reached the same conclusion. Although the model fitness variability was comparatively large when the random sample size was small, when we analyzed data from five randomly chosen participants over 10,000 iterations, the best-fit model changed only once, without a significant difference

from the second-best model. Thus, we empirically demonstrated the robustness of data obtained through a crowdsourced model-comparison experiment, at least for our task of bivariate pointing. This work will contribute to research on novel GUI operation models, and it will motivate us to conduct further studies on exploring other applicable tasks.

References

1. Accot, J., Zhai, S.: Beyond fitts' law: models for trajectory-based hci tasks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '97). pp. 295–302 (1997). https://doi.org/10.1145/258549.258760
2. Accot, J., Zhai, S.: Refining fitts' law models for bivariate pointing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 193–200. CHI '03, ACM, New York, NY, USA (2003). https://doi.org/10.1145/642611.642646, http://doi.acm.org/10.1145/642611.642646
3. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6), 716–723 (1974). https://doi.org/10.1109/TAC.1974.1100705
4. Appert, C., Chapuis, O., Beaudouin-Lafon, M.: Evaluation of pointing performance on screen edges. In: Proceedings of the Working Conference on Advanced Visual Interfaces. pp. 119–126. AVI '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1385569.1385590, http://doi.acm.org/10.1145/1385569.1385590
5. Bi, X., Li, Y., Zhai, S.: Ffitts law: Modeling finger touch with fitts' law. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1363–1372. CHI '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2470654.2466180, http://doi.acm.org/10.1145/2470654.2466180
6. Bohan, M., Longstaff, M., Van Gemmert, A., Rand, M., Stelmach, G.: Effects of target height and width on 2d pointing movement duration and kinematics. Motor control **7**, 278–289 (08 2003)
7. Burnham, K.P., Anderson, D.R.: Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media (2003)
8. Cockburn, A., Lewis, B., Quinn, P., Gutwin, C.: Framing effects influence interface feature decisions. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–11. CHI '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3313831.3376496, https://doi.org/10.1145/3313831.3376496
9. Crossman, E.R.: The measurement of perceptual load in manual operations. Ph.D. thesis, University of Birmingham (1956)
10. Devore, J.L.: Probability and Statistics for Engineering and the Sciences. Brooks/Cole, 8th edn. (January 2011), iSBN-13: 978-0-538-73352-6
11. Faridani, S.: Models and Algorithms for Crowdsourcing Discovery. Ph.D. thesis, USA (2012)
12. Findlater, L., Zhang, J., Froehlich, J.E., Moffatt, K.: 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. pp. 6813–6824. CHI '17, ACM,

New York, NY, USA (2017). https://doi.org/10.1145/3025453.3025820, http://doi.acm.org/10.1145/3025453.3025820
13. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology **47**(6), 381–391 (1954). https://doi.org/10.1037/h0055392
14. Gan, K.C., Hoffmann, E.R.: Geometrical conditions for ballistic and visually controlled movements. Ergonomics **31**(5), 829–839 (1988). https://doi.org/10.1080/00140138808966724
15. Goldberg, K.Y., Faridani, S., Alterovitz, R.: Two large open-access datasets for fitts' law of human motion and a succinct derivation of the square-root variant. IEEE Transactions on Human-Machine Systems **45**(1), 62–73 (2015). https://doi.org/10.1109/THMS.2014.2360281
16. Gori, J., Rioul, O., Guiard, Y.: Speed-accuracy tradeoff: A formal information-theoretic transmission scheme (fitts). ACM Trans. Comput.-Hum. Interact. **25**(5) (Sep 2018). https://doi.org/10.1145/3231595, https://doi.org/10.1145/3231595
17. Gould, S.J.J., Cox, A.L., Brumby, D.P.: Diminished control in crowdsourcing: An investigation of crowdworker multitasking behavior. ACM Trans. Comput.-Hum. Interact. **23**(3) (Jun 2016). https://doi.org/10.1145/2928269, https://doi.org/10.1145/2928269
18. Hoffmann, E.R., Drury, C.G., Romanowski, C.J.: Performance in one-, two- and three-dimensional terminal aiming tasks. Ergonomics **54**(12), 1175–1185 (2011). https://doi.org/10.1080/00140139.2011.614356, https://doi.org/10.1080/00140139.2011.614356
19. Hoffmann, E.R., Sheikh, I.H.: Effect of varying target height in a fitts' movement task. Ergonomics **37**(6), 1071–1088 (1994). https://doi.org/10.1080/00140139408963719
20. Ko, Y.J., Zhao, H., Kim, Y., Ramakrishnan, I., Zhai, S., Bi, X.: Modeling two dimensional touch pointing. In: Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. pp. 858–868. UIST '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3379337.3415871, https://doi.org/10.1145/3379337.3415871
21. Komarov, S., Reinecke, K., Gajos, K.Z.: Crowdsourcing performance evaluations of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 207–216. CHI '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2470654.2470684, http://doi.acm.org/10.1145/2470654.2470684
22. MacKenzie, I.S.: Fitts' law as a research and design tool in human-computer interaction. Human-Computer Interaction **7**(1), 91–139 (1992). https://doi.org/10.1207/s15327051hci0701_3
23. MacKenzie, I.S., Buxton, W.: Extending fitts' law to two-dimensional tasks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 219–226. CHI '92, ACM, New York, NY, USA (1992). https://doi.org/10.1145/142750.142794, http://doi.acm.org/10.1145/142750.142794
24. MacKenzie, I.S., Isokoski, P.: Fitts' throughput and the speed-accuracy tradeoff. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1633–1636. CHI '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1357054.1357308, https://doi.org/10.1145/1357054.1357308

25. Matejka, J., Glueck, M., Grossman, T., Fitzmaurice, G.: 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. pp. 5421–5432. CHI '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858063, https://doi.org/10.1145/2858036.2858063
26. Meyer, D.E., Abrams, R.A., Kornblum, S., Wright, C.E., Keith Smith, J.E.: Optimality in human motor performance: ideal control of rapid aimed movements. Psychological review **95**(3), 340–370 (July 1988). https://doi.org/10.1037/0033-295x.95.3.340
27. Rioul, O., Guiard, Y.: Power vs. logarithmic model of fitts' law: A mathematical analysis. Mathematical Social Sciences **2012**, 85–96 (12 2012). https://doi.org/10.4000/msh.12317
28. Schwab, M., Hao, S., Vitek, O., Tompkin, J., Huang, J., Borkin, M.A.: Evaluating pan and zoom timelines and sliders. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12. CHI '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300786, https://doi.org/10.1145/3290605.3300786
29. Senanayake, R., Hoffmann, E.R., Goonetilleke, R.S.: A model for combined targeting and tracking tasks in computer applications. Experimental Brain Research **231**(3), 367–379 (Nov 2013). https://doi.org/10.1007/s00221-013-3700-4
30. Soukoreff, R.W., MacKenzie, I.S.: 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 (2004). https://doi.org/10.1016/j.ijhcs.2004.09.001
31. Wobbrock, J.O., Findlater, L., Gergle, D., Higgins, J.J.: The aligned rank transform for nonparametric factorial analyses using only anova procedures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 143–146. CHI '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/1978942.1978963, http://doi.acm.org/10.1145/1978942.1978963
32. Yamanaka, S.: Steering performance with error-accepting delays. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 570:1–570:9. CHI '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300800, http://doi.acm.org/10.1145/3290605.3300800
33. Yamanaka, S., Miyashita, H.: Modeling the steering time difference between narrowing and widening tunnels. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 1846–1856. CHI '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858037, http://doi.acm.org/10.1145/2858036.2858037
34. Yamanaka, S., Shimono, H., Miyashita, H.: 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. pp. 157–169. ISS '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3343055.3359698
35. Yamanaka, S., Usuba, H.: Calibration methods of touch-point ambiguity for finger-fitts law (2021), https://arxiv.org/abs/2101.05244

36. Yang, H., Xu, X.: Bias towards regular configuration in 2d pointing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1391–1400. CHI '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1753326.1753536, http://doi.acm.org/10.1145/1753326.1753536
37. Zhai, S., Kong, J., Ren, X.: 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 (2004). https://doi.org/10.1016/j.ijhcs.2004.09.007
38. Zhang, X., Zha, H., Feng, W.: Extending fitts' law to account for the effects of movement direction on 2d pointing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 3185–3194. CHI '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2207676.2208737, http://doi.acm.org/10.1145/2207676.2208737
39. Zhao, J., Soukoreff, R.W., Ren, X., Balakrishnan, R.: A model of scrolling on touch-sensitive displays. International Journal of Human-Computer Studies **72**(12), 805 – 821 (2014). https://doi.org/https://doi.org/10.1016/j.ijhcs.2014.07.003, http://www.sciencedirect.com/science/article/pii/S1071581914000998