Time-Penalty Impact on Effective Index of Difficulty and Throughputs in Pointing Tasks

Shota Yamanaka, Keisuke Yokota, Takanori Komatsu

Time-Penalty Impact on Effective Index of Difficulty and Throughputs in Pointing Tasks

Shota Yamanaka[1], Keisuke Yokota[2] and Takanori Komatsu[2]

[1] Yahoo Japan Corporation (syamanak@yahoo-corp.jp)
[2] Meiji University

Abstract. In realistic graphical user interfaces, clicking outside a target would require recovery time from an error, e.g., selecting an unintended hyperlink requires reloading the previous webpage. Several studies on target-pointing tasks have examined the effects of a "penalty time" for mis-clicks on movement time and error rate, but the effects on throughput (i.e., a unified metric on pointing performance) have not been thoroughly investigated. We conducted a crowdsourcing study with 127 workers and a lab-based controlled study with 30 university students. The penalty times varied from 0 to 10 sec, and the results consistently indicated that the throughput differences were less than 5%, although the error rates were remarkably different when the penalty time was 0 sec. This demonstrated the potential of normalization capability of the effective width method of Fitts' law, and the throughput is considered a valid metric when researchers would like to compare several task conditions in realistic user interfaces in which error operations induce different recovery times. However, because the model fitness using the effective width method was comparatively low when the penalty time was 0 sec, comparing throughputs for different conditions is not recommended. We also discussed potential issues related to the effective width method of Fitts' law, such as endpoints not following a linear relationship to the given target width only under the zero-penalty condition.

Keywords: Performance modeling · Fitts' law · Crowdsourcing.

1 Introduction

We investigated the effects of imposing a penalty time for an operation detected as incorrect on user performance in target-pointing tasks. In typical pointing tasks, clicking on a target is considered a success, and missing it is considered an error. In the latter case, participants would be asked to click on the target again until success (e.g., [44]) or the experimental system ends the current trial and proceeds to the next (e.g., [6]). However, in realistic graphical user interfaces (GUIs), if users miss a target, they would have to perform additional operations to recover from the error. For example, if a user misses an intended hyperlink and the neighboring link is clicked, they must go back to the previous webpage, wait for the page to reload, then try to click the intended link again.

Such a recovery time from an error changes depending on the networking status, applications, and so on. It may take 3–5 sec to recover after clicking on an unintended link, while clicking on the surrounding empty space incur no additional recovery time (i.e., only a retry is needed). For menu selection, e.g., in Microsoft Paint, mis-clicking on the top-level items (File / Home / View) incurs no recovery time, but if users mis-click the final item in "File → Save as → Export → JPEG picture," a longer retry time is incurred and additional effort is needed. In these realistic tasks, mis-operations and the resulting fatigue will affect user performance. As a more serious case, if users accidentally click on the "Pay" button on the PayPal site, or click the "Buy now with 1-Click" button on the Amazon site, a cancelling email has to be sent or the credit card company must be told to stop the payment, which may take several minutes.

To investigate the effects of such a recovery time on user performance, imposing a penalty time for mis-clicking in target pointing tasks has been examined [3, 15, 41]. For example, in Banovic et al.'s study, the participants had to wait for a certain time before the task could be resumed [3]. They theoretically and empirically determined that the error rate decreases and the cursor-movement time increases as the penalty time is increased in a mouse-pointing task [3]. This is intuitive because participants would obviously try to point to the target more carefully in a longer penalty time to shorten the overall task-completion time.

However, Banovic et al. reported three separate indicators of user performance: the time for the first click, time for task completion including the penalty, and error rate. For target-pointing studies, it is known that reporting a unified metric called *throughput*, which is used to measure user performance after normalizing the error rate, is needed to compare devices and techniques [23, 26, 31, 35, 43]. To compare two or more experimental conditions, such as comparing input devices (e.g., [9]) or user groups (e.g., children vs. older adults [32]), the trade-off between speed and accuracy typically differs under each condition, thus normalizing the error rates if needed. Therefore, when researchers would like to determine if, e.g., "there is no significant difference in user performance for 1- and 10-sec penalty times," using the throughputs is preferred.

In this paper, we empirically explore how the penalty time affects throughputs from two viewpoints. First, it has never been studied if the effective width method of Fitts' law [11] holds for pointing tasks with penalty time. Because the throughput is valid if the task is modeled with this method, we have to examine the model fitness. Second, the normalization capability of throughput is unclear. Theoretically, even if participants are biased towards either speed or accuracy, throughputs should not change [28]. In our case, even when we change the penalty time from 0 to 10 sec, while the first click time and the error rate could change, the throughputs should not significantly differ. If we confirm these two facts (Fitts' law holds and throughputs are not significantly different), researchers and designers will be able to compare different devices and user groups in realistic GUIs requiring recovery times, which contributes to future human-computer interaction (HCI) studies and GUI design.

We conducted two user experiments: crowdsourcing involving 127 workers and a conventional lab-based study involving 30 university students. Because Fitts' law and the effective width method are for modeling the central tendency of user performance [35], crowdsourcing is useful for recruiting many participants. However, there are issues with crowdsourcing GUI experiments, e.g., some workers may not follow given instructions [12]. Therefore, we were concerned if some of the workers performed the pointing task but ignored the instruction of "Minimize the whole task-completion time including the pointing time and penalty time." Hence, we also conducted the lab-based experiment, which was a self-replication study. Because these two experiments had different advantages, we do not need to compare the results directly; for example, a comparison such as "Crowdworkers were significantly faster than lab-based participants ($p < 0.05$)" is not necessary. Our main findings are as follows:

- For the effective width method of Fitts' law, the model fitness under the zero-penalty condition was remarkably low ($R^2 = 0.232$ and 0.601 for the crowdsourcing and lab-based experiments, respectively), while the other penalty conditions showed $R^2 > 0.9$. This low fit under the zero-penalty condition is consistent with previous studies [37, 44]. This questions the reliability of throughput data under the zero-penalty condition.
- The throughputs for the 11 penalty time conditions were close: within 3 and 4% for the crowdsourcing and lab-based experiments, respectively, while the error rates more clearly differed depending on the penalty times. This demonstrated the normalization capability of throughput.

2 Related Work

2.1 Fitts' Law and the Effective Width Method

According to Fitts' law, the time for the first click, or movement time MT, to point to a target relates to the index of difficulty ID in bits [13]:

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

where a and b are empirical regression constants. The Shannon formulation of ID [26] is widely used in HCI:

$$ID_n = \log_2 (A/W + 1), \tag{2}$$

where the target distance (or amplitude) is A and its width is W. This ID is the *nominal* value using nominal A and W drawn on the display.

Typically, participants are asked to point to a target as quickly and accurately as possible [35], but some participants tend to show short MTs and high error rates (ERs), while others show long MTs and low ERs [44]. Thus, when researchers compare the performances of several devices or several user groups, normalizing ERs is necessary. Using Crossman's post-hoc correction for W [11] is recommended in HCI [35] and the ISO standard [23].

$$W_e = 4.133 \cdot SD_x, \tag{3}$$

where W_e is the effective width and SD_x is the standard deviation of the actual click positions (or *endpoints*) along the movement axis. We then obtain the effective index of difficulty ID_e by replacing the W in Equation 2 with W_e.

Researchers can use a unified measure of user performance called throughput [bits/sec] that integrates the speed (MT) and ID_e [35]:

$$\text{Throughput} = \frac{1}{|A| \times |W|} \sum_{i=1}^{|A|\times|W|} \left(\frac{ID_{e_i}}{MT_i} \right), \tag{4}$$

where i indicates the i-th condition among $|A|\times|W|$. We calculate the throughput for a participant then compute the grand throughput by averaging all participants' throughputs [35][3].

The basis of this adjustment is that a spread of hits follows a normal distribution over a target. Using this method, W_e is adjusted so that 3.88% ($\sim$4%) of clicks fall outside the target; thus, we can compare the throughputs, e.g., from different user groups. Although the effective width method has issues particularly with its theoretical bases [16], its empirical benefits have been recognized [44].

2.2 Penalty-Time Paradigm

Previous Work on Target-Pointing Task. It has been demonstrated that, regardless of whether a participant's priority biases towards speed or accuracy, Fitts' law using ID_n holds [14, 44]. While these biases have been controlled by monetary incentive/penalty [14] or by oral instruction [44], Banovic et al. determined that the bias is affected by the risk associated with the penalty time (0–20 sec) for target misses [3]. Gillan et al. compared 0- and 30-sec penalty conditions with a mouse [15], and the latter showed longer MT and lower ER, which was consistent with Banovic et al.'s report.

These studies consistently reported that, if the penalty time for a target miss was long (e.g., 10 sec), the participants would attempt to more carefully point to the target. While this strategy lengthens the MT for a single click, the overall task-completion time should be shorter than if the participant performs the task quickly and inaccurately. When the penalty time is short (e.g., 1 sec), the risk of the overall task-completion time being lengthened is not serious, even if a participant rapidly aims for the target and misses it in several trials. This strategy may reduce the MT per click, so the overall task-completion time would be shorter than with the slow-and-careful strategy. Thus, participants implicitly balance (optimize) the speed-accuracy trade-off for a given penalty time [3]. However, as Banovic et al. found, while a longer penalty time monotonically increases the single-task-completion time, the effect quickly plateaus (see Figure 4 in [3]).

[3] Olafsdottir et al. listed 20 approaches to compute throughput [31]. We used Soukoreff and MacKenzie's method [35].

Banovic et al. showed that the difference in single-task-completion times under 3- and 30-sec penalty conditions was less than 0.1 sec, regardless of the Fitts' law difficulty. Moreover, the *ER*s for 3.33- and 6.67-sec penalties were not significantly different, whereas they were significantly worse for a 0-sec penalty. Since the effects of penalty time on task completion times are assumed to level off quickly, using an extremely long penalty time, such as 30 sec, was unnecessary in our experiments, although it would occur in realistic GUIs, as discussed in the Introduction.

Inspired by the work of Banovic et al., Yamanaka evaluated the effects of penalty times (0–4 sec) in touch-pointing tasks with 2D square targets on tablets [39]. The results indicated that the task-completion time was not significantly affected by penalty time, whereas *ER* significantly decreased as penalty time increased. This partially reproduced Banovic et al.'s findings [3]. Furthermore, Yamanaka et al. had crowdworkers perform almost the same task with 1D and 2D targets [41], but the penalty time was fixed at 3 sec.

Unexplored Space for Penalty-Time Paradigm in Pointing Task. In summary of the previous studies on penalty time, the speed-accuracy trade-off can be controlled by changing the penalty time [3, 15, 39], which have been confirmed in lab-based experiments. A limitation in these studies is that they reported the results on times and errors separately; thus, the effect of penalty time on throughput remains unclear. Our study bridges the gap between previous studies on penalty time [3, 15, 39] and the standardized methodology of Fitts' law that unifies the speed (MT) and accuracy (SD_x) into a single throughput [23, 35].

2.3 Crowdsourcing User Studies Compared with Lab-based Ones

Research fields other than HCI have found high internal validity of crowdsourced data in, for example, political science [4] and behavioral economics [21] experiments. For user experiments on GUI operations such as target pointing and menu selection, Komarov et al. concluded that crowdsourcing can be used for conducting performance evaluations of GUIs and that it is a complementary approach to lab-based experiments, with greater diversity of participants and less effort for recruitment [25].

In contrast, by using more powerful statistical analysis methods and recruiting many more participants for lab-based experiments, Findlater et al. showed that crowdworkers have significantly shorter MTs and higher *ER*s both in mouse- and touch-pointing tasks [12]. They reported a good model fitness of Fitts' law ($r = 0.926$ using mice). In addition, Schwab et al. showed that a crowdsourcing scrolling task on a desktop environment resulted in a Fitts' law fitness of $R^2 = 0.983$ [33]. These results of good Fitts' law fitness motivated us to conduct an experiment for model fitness and performance evaluation on a crowdsourcing platform.

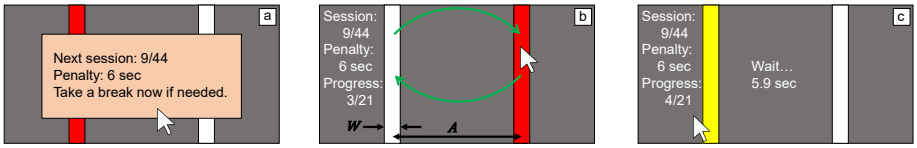

Fig. 1. Abstracted image of the experimental system. (a) The penalty time is shown before each session. (b) Serial-target-pointing task in which a worker attempts to click alternating red targets. (c) If a worker misses the target, the target turns yellow, and the cursor cannot be moved for the specified penalty time.

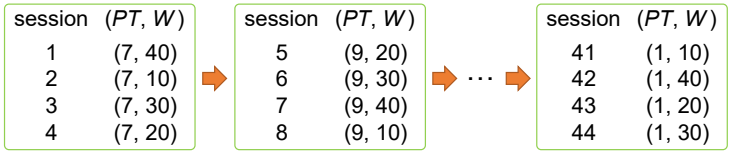

Fig. 2. Example case of 44 sessions and the random order of PT and W

3 Experiment 1: Crowdsourcing User Study

We conducted a 1D serial-target-pointing experiment (Figure 1) on the *Yahoo! Crowdsourcing* platform (https://crowdsourcing.yahoo.co.jp) from September 9 to 11, 2020. The task was offered only in Japan. The study was approved through our company's IRB-equivalent research ethics team. They raised no specific concerns or requested any changes.

3.1 Task and Procedure

The experimental system was developed with the `Processing` language (version 3.5.4). The crowdworkers downloaded the executable file from a given URL and ran it. They first completed a questionnaire on their age (numeric), gender (free form to allow for a non-binary or arbitrary answer), handedness (left or right), input device (free form), and history of PC use (numeric in years). The system then proceeded to the pointing-task phase.

There were 11 penalty times (PTs) and 4 Ws. They were fully crossed with each other, so each worker completed 44 *sessions* in total (see Figure 2). Each session consisted of 21 cyclic clicks back and forth between the left and right targets. Before the next session began, the number out of the 44 sessions and PT were displayed, along with a message about taking a break (see Figure 1a). Clicking on the message area initiated the session.

In the pointing-task phase, the workers were asked to click on the red vertical bar. If the worker clicked on the bar, the colors of target and non-target bars (red and white, respectively) changed, as shown in Figure 1b. If the worker pressed the mouse button when the cursor was outside the target, we call this "a misclicked position." In this case, the target bar's color turned yellow, and the worker

could not move the cursor from the mis-clicked position until the PT expired (Figure 1c). Technically, the system moved the cursor to the mis-clicked position every frame (60 fps by default) by using the `java.awt.Robot.mouseMove` function, and the target did not sense the mouse-click event. When the PT expired, the cursor began moving from the mis-clicked position; i.e., not retrying a new trial.

While Banovic et al. stopped the display of the cursor during the PT [3], we were concerned that some workers might think that there was a bug in the program if the cursor disappeared, so we continued displaying it. Because we could not be sure that all the workers would be able to hear sounds during the task, we did not give auditory feedback for success or failure. The remaining PT was displayed as a countdown timer.

3.2 Design

The experiment was an 11×4 within-subjects design with the following independent variables and levels: $PT = 0$ to 10 sec in 1-sec steps and $W = 10$, 20, 30, and 40 pixels. While Banovic et al. tested $PT = 0$, 3.33, 6.67, 10, and 20 sec, they confirmed that the effects of PT on task-completion times and ERs quickly plateau [3], and thus we prioritized testing a shorter range of PTs precisely with 1-sec steps. Because error-recovery times vary even for a single application or website, we decided to use a within-subjects design for PT. The choice of W is independent from the throughput analysis [28, 35] but affects ER [13]; thus, we used somewhat narrow targets. We denote the levels with subscripts, e.g., "W_{20}."

In each session, the first target was on the left side. To measure the central tendency of each worker's performance under each $PT \times W$ condition, requiring 15 to 25 clicks is recommended [35], so we considered the first 5 clicks to be practice and used the remaining 16 clicks (8 clicks for each side) for data collection. In each session, PT and W were fixed. For four successive sessions, PT was fixed, and the order of the four W conditions was randomized (Figure 2).

The ER for target pointing is assumed to depend solely on W; the target distance A is not an important factor because it does not strongly affect the click-point distribution [5, 22, 44]. Hence, we fixed A and varied PT widely and precisely, which was the focus of this study. Using a single A is common for studies on measuring throughput [28, 31], and since Fitts' law analysis applies even if only the target size changes (e.g., [11, 20]), A was fixed at 600 pixels. We recorded a total of $11_{PT} \times 4_W \times 16_{\text{repetitions}} \times 127_{\text{workers}} = 89{,}408$ data points.

3.3 Participants

A total of 127 workers completed the task. Their demographics were as follows. Age: ranging from 23 to 64 years, $M = 43.1$ and $SD = 8.39$. Gender: 104 were male and 23 were female. Handedness: 6 were left-handed and 121 were right-handed. Input devices: 6 used a touchpad, 1 a trackball, 1 a trackpoint, 1 a pen

tablet (whether direct or indirect stylus input is unknown), and 118 a mouse. PC usage history: from 5 to 40 years, $M = 21.0$ and $SD = 6.49$.

We recruited workers who used Windows (Vista or a later version) because our system runs in only those environments. No other qualifications, such as the worker's skills, were required. Once workers accepted the task, they were asked to read the online instructions explaining the task. The instructions stated that they should complete the task in as a short a time as possible including the PT [3, 38, 39, 41]. They also explained that a PT would be imposed if they missed the target. Because we wanted them to understand how the cursor would remain fixed if they missed a target, we asked them to watch a short video in which one of the authors had performed the task and missed the target.

After a worker finished all 44 sessions, the log data, which included the questionnaire data, clicked positions, and timestamps, were exported to a csv file. The workers uploaded the file to a server to receive payment. The task typically took 20 min, and the payment was JPY 100 (∼USD 0.96), so the effective hourly payment was about JPY 300. This amount was determined after a discussion with the platform's advisor; although it was less than that on other platforms such as Amazon Mechanical Turk, it was common for that platform. Previous work has shown that the payment affects workers' motivation [7], and thus our conclusion could change if we set a much higher payment.

3.4 Results

Screening Outlier Data For Fitts' law tasks, there are two types of trial-level outliers: spatial and temporal. In addition, there are participant-level outliers in the temporal results: data for workers who performed extremely slowly or rapidly were removed from the analysis. These outliers are described as follows.

A spatial outlier was a position clicked more than 3σ from the mean clicked position. A temporal outlier was a trial in which the time taken to make the first click, i.e., MT, was more than 3σ from the mean MT. These trial-level outlier calculations were run for each session for each worker. This 3σ criterion was used by MacKenzie and colleagues [29, 27, 35].

For the participant-level outliers, we calculated the mean MT across all 44 conditions ($11_{PT} \times 4_W$) for each participant. The data for workers whose mean MT was more than 3σ from the average MT across all workers were removed as outliers. We detected the trial- and participant-level outliers independently.

There are various other approaches to detecting outliers, such as the Smirnov-Grubbs' test and inter-quartile range method, as well as the Fitts' law-specific method, i.e., a cursor's movement distance was less than $A/2$ or the clicked position was more than $2W$ from the target center [3, 12]. Although investigating a more robust data-screening method is an important topic in crowdsourcing research, discussing this point by analyzing the data with several criteria is clearly beyond the scope of this study. Hence, in our data analysis, we used a single method (the 3σ criterion).

We found 55 spatial and 378 temporal trial-level outliers (0.5% of the data). Two participant-level outlier workers were eliminated due to longer MTs than

the mean. Notably, while the average MT was approximately 1 sec, one of the outlier workers had a MT longer than 10 sec twice (60 and 307 sec), which indicates an obvious lack of concentration. Integrating the trial- and participant-level outliers led to the removal of 2.02% of the data points, which is close to the rate in a previous study [12].

Analyses of Dependent Variables After the outliers were removed, 87,602 data points (98.0%) were analyzed. As with typical pointing studies using the effective width method, the dependent variables were MT, ER, and throughput. MT data are typically distributed normally, but the Shapiro-Wilk test ($\alpha = 0.05$) showed that 31 of the 44 conditions violated the normality assumption, so we log-transformed the data before applying repeated-measures ANOVA with Bonferroni's p-value adjustment method for pairwise comparisons. ERs are nonparametric data, so we used ANOVAs with *Aligned Rank Transform* [36] and Tukey's p-value adjustment method for pairwise comparisons. Because throughput depends on both MT and click point distribution (SD_x), we cannot interpret this as parametric data, so we again used ANOVAs with Aligned Rank Transform. Because the throughput merged the four Ws (Equation 4), the only independent variable was PT.

To simplify the result statements, we mainly report on the effects of PT and briefly report those of W on the dependent variables. For example, we avoid reporting the results for all combinations of significantly different PT pairs in the $PT \times W$ interaction among the 220 possible pairs (${}_{11}\mathrm{C}_2 \times {}_4\mathrm{C}_1$).

Movement Time For the F statistic, the degrees of freedom for the main effects of PT and W, as well as the interaction of $PT \times W$, were corrected using the Greenhouse-Geisser method because Mauchly's sphericity assumption was violated ($\alpha = 0.05$). We found significant main effects of PT ($F_{3.784,469.2} = 9.897$, $p < 0.001$, $\eta_p^2 = 0.74$) and W ($F_{1.441,180.2} = 2280$, $p < 0.001$, $\eta_p^2 = 0.95$). The interaction of $PT \times W$ was significant ($F_{20.12,2495} = 1.544$, $p < 0.001$, $\eta_p^2 = 0.57$). As expected, the MTs decreased as W increased ($p < 0.001$ for all pairs). The mean MTs for W_{10} to W_{40} were 1.31, 1.05, 0.939, and 0.874 sec, respectively.

The mean MTs tended to increase as PT increased, as shown in Figure 3a, which means that the workers became more careful. Although there was a number of significantly different pairs, the mean MTs ranged from 1.00 to 1.06 sec. This slight effect of PT on MT can also be confirmed with a linear regression. A model of $MT = a + b \cdot ID_n + c \cdot PT$ yields adjusted $R^2 = 0.988$ and $c = 0.00439$ (with $p < 0.0001$). This means that even when PT increases from 0 to 10 s, MT is assumed to increase by only 0.04 sec.

Error Rate We found significant main effects of PT ($F_{10,1240} = 25.87$, $p < 0.001$, $\eta_p^2 = 0.17$) and W ($F_{3,372} = 156.7$, $p < 0.001$, $\eta_p^2 = 0.56$). The interaction of $PT \times W$ was significant ($F_{30,3720} = 13.59$, $p < 0.001$, $\eta_p^2 = 0.099$). As expected, the ER decreased as W increased ($p < 0.001$ for all pairs). The mean ERs for W_{10} to W_{40} were 2.96, 2.25, 1.91, and 1.40%, respectively.

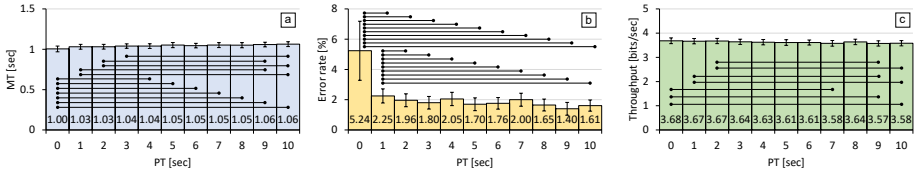

Fig. 3. Main effects of PT on (a) MT without log-transformation, (b) ER, and (c) throughput in Experiment 1. Error bars show 95% CIs across all 125 non-outlier workers, and horizontal lines indicate significantly different pairs (at least $p < 0.05$).

Table 1. Model fitness in Experiment 1. The yellow cell shows the lowest fit using ID_e. The pink cell indicates an insignificant ($p > 0.05$) contributor of ID_e to predict MTs. The blue cells indicate that there were significant differences in fits using ID_n and ID_e.

PT	Nominal (ID_n)				Effective (ID_e)				Difference		
[sec]	R^2	p value	AIC	BIC	R^2	p value	AIC	BIC	R^2	AIC	BIC
0	0.996	0.00203	-21.7	-22.9	0.232	0.518	-0.715	-1.94	0.764	-21.0	-21.0
1	0.984	0.00796	-15.6	-16.8	0.920	0.0410	-9.09	-10.3	0.0641	-6.49	-6.51
2	0.994	0.00284	-20.0	-21.2	0.993	0.00358	-19.1	-20.3	0.00132	-0.904	-0.932
3	0.988	0.00598	-16.7	-17.9	0.975	0.0125	-13.7	-15.0	0.0131	-2.97	-2.90
4	0.990	0.00493	-17.5	-18.7	0.998	0.00116	-23.3	-24.5	-0.00784	5.81	5.78
5	0.987	0.00652	-16.1	-17.4	0.980	0.0103	-14.3	-15.6	0.00700	-1.85	-1.77
6	0.992	0.00400	-18.3	-19.5	0.981	0.00932	-14.9	-16.1	0.0110	-3.37	-3.40
7	0.989	0.00553	-16.9	-18.1	0.980	0.00990	-14.6	-15.8	0.00896	-2.28	-2.31
8	0.993	0.00343	-18.9	-20.1	0.943	0.0291	-10.4	-11.6	0.0502	-8.47	-8.50
9	0.993	0.00355	-18.9	-20.1	0.998	0.00103	-23.8	-25.1	-0.00508	4.91	4.98
10	0.989	0.00560	-16.3	-17.5	0.973	0.0134	-12.8	-14.1	0.0158	-3.52	-3.45

The ER for PT_0 was remarkably high, as shown in Figure 3b. The ER for PT_1 was nevertheless significantly different from those of $PT \geq 2$ sec. Finally, there were no significant differences between any pair for $PT \geq 2$ sec.

Throughput and Fitts' Law Fitness We found a significant main effect of PT ($F_{10,1240} = 4.221$, $p < 0.001$, $\eta_p^2 = 0.033$). Pairwise comparisons showed that the throughputs tended to decrease as PT increased, as shown in Figure 3c. The highest throughput was observed for PT_0, and the lowest was for PT_9. However, the difference was only $3.68 - 3.57 = 0.11$ bits/sec (i.e., 3% of the baseline condition, PT_0).

This indicates that throughput performance is not strongly affected by PT, which is also supported by the small effect size of PT ($\eta_p^2 = 0.033$). However, Figure 3c is somewhat misleading. According to the Fitts' law fitness using ID_e, the correlation for PT_0 was remarkably low ($R^2 = 0.232$, see the yellow cell in Table 1) while it was > 0.9 under the other PT conditions. Because judging model fitness on the basis of only the absolute R^2 is problematic [17], we argue that Fitts' law cannot capture the PT_0 condition on the basis of the p value for ID_e. As shown with the pink cell in Table 1, ID_e was not a significant

Table 2. Results of Experiment 1. The blue cells indicate a violation of Fitts' law's assumption that W and SD_x are linearly related.

PT	MT [sec]				W_e [pixels]				R^2	ER	Throughput
[sec]	W_{10}	W_{20}	W_{30}	W_{40}	W_{10}	W_{20}	W_{30}	W_{40}	(W, SD_x)	[%]	[bits/sec]
0	1.25	1.03	0.910	0.857	78.5	66.5	86.9	91.0	0.477	5.24	3.68
1	1.30	1.02	0.928	0.865	32.1	46.7	75.2	89.7	0.981	2.25	3.67
2	1.28	1.04	0.929	0.866	24.1	46.6	68.5	89.8	1.000	1.96	3.67
3	1.32	1.04	0.942	0.879	32.0	52.9	67.7	89.6	0.995	1.80	3.64
4	1.30	1.04	0.932	0.869	24.1	52.1	76.3	96.9	0.995	2.05	3.63
5	1.33	1.05	0.943	0.884	32.7	53.3	75.2	90.4	0.995	1.70	3.61
6	1.31	1.05	0.951	0.870	23.6	46.6	76.9	89.6	0.978	1.76	3.61
7	1.32	1.05	0.943	0.883	32.2	54.9	75.4	97.4	1.000	2.00	3.58
8	1.32	1.06	0.959	0.877	30.5	44.9	68.4	88.3	0.992	1.65	3.64
9	1.32	1.07	0.954	0.892	30.4	54.6	74.6	89.4	0.989	1.40	3.57
10	1.36	1.06	0.957	0.882	23.3	45.3	68.0	98.7	0.993	1.61	3.58

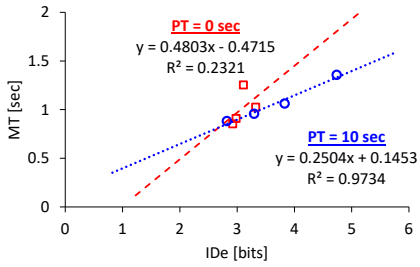

Fig. 4. Fitts' law fitness using ID_e for PT_0 (red) and PT_{10} (blue).

contributor to explain the MT only for PT_0 ($p = 0.518 > 0.05$). Thus, predicting worker performance under the zero-penalty condition on the basis of throughput is not reliable because throughput is an indicator of performance only when the operation can be modeled by Fitts' law. To the best of our knowledge, this finding, i.e., that Fitts' law fitness using the effective width method is notably low ($R^2 = 0.232$) for crowdsourced data, is a novel empirical finding that cautions against measuring worker performance on the basis of only throughput.

In contrast, using ID_n showed $R^2 > 0.98$. To statistically compare the model-fitness difference, we used AIC [2] and BIC [24]. For simplicity, we consider an AIC difference greater than 10 to be significant [8]. This was also applied to BIC [24]. The results indicate that, only under the PT_0 condition, the model fitness using ID_e is significantly inferior to ID_n (see blue cells in Table 1). This comparatively low model fitness using ID_e is consistent with previous studies [37, 44]. This result again supports the low reliability of throughput analysis under the PT_0 condition, while predicting MT using ID_n is not problematic.

We examine this low fitness only for PT_0 using ID_e in more detail by visualizing the fits. As example cases, we plot the data for PT_0 and PT_{10} in Figure 4. While the four data points for PT_{10} are close to the regression line, those for PT_0

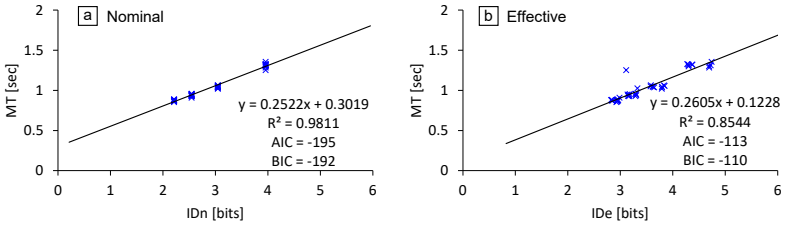

Fig. 5. Model fitness using (a) ID_n and (b) ID_e for 44 data points in Experiment 1.

are not. More critically, the spread of the data points on the x-axis for PT_0 is narrower than that for PT_{10}. This is because the distributions of the click positions (SD_x values) fell in a narrow range. In particular, for W_{10}, $W_e = 4.133 \cdot SD_x$ was greater than that for W_{20} (see blue cells in Table 2). This clearly shows that, for W_{10}, W_e under the PT_0 condition was large compared with the other PT conditions. This was likely due to workers "giving up" pointing to the smallest target accurately on the first attempt, as there was no penalty for mis-clicks. This is supported by the highest ER (8.05%) observed under the $PT_0 \times W_{10}$ condition.

This result violates the assumption of Fitts' law, especially for the effective width method, because this method is based on the fact that the endpoint variability is linearly (or proportionally) related to the given W: $SD_x = a + b \cdot W$ [5, 11, 26, 35]. The results from the conditions other than PT_0 validate this assumption, with R^2 between W and SD_x being over 0.97 (see Table 2). More specifically, for PT_0, the larger W_e for W_{10} than that for W_{20} results in smaller ID_e for W_{10} than that for W_{20}, meaning "pointing to a narrower target is easier," which is obviously incorrect.

Normalization Capability of the Effective Width Method Even if ERs differ under several task conditions (here, the 11 PTs), using ID_e normalizes the ERs; thus, the Fitts' law fitness without separating the task conditions would show a higher fit compared with ID_n. This benefit was empirically demonstrated by Zhai et al. [44]. However, as shown in Figure 5, we did not confirm such a capability. Because the change in MT due to the 11 PTs was small (0.06 sec at most), we did not see the benefit of using ID_e for comparing different conditions.

4 Experiment 2: Lab-based User Study

This experiment was conducted in a silent room of our university from December 3 to 16, 2020. We followed our university's compliance policy for in-person user experiments regarding Covid-19. In particular, there should be no more than two individuals in a room at the same time: in our study there were always an experimenter and a participant. We also followed the other requirements, e.g., air ventilation, sanitation of equipment, and mask mandate.

We used the same experimental system with the same task design as in Experiment 1. The procedure, e.g., watching the instruction video and filling the questionnaire, was also the same. The only difference was that we asked the participants to use the apparatus we prepared.

4.1 Apparatus

The PC that we used was a Microsoft Surface Laptop 3 (AMD Ryzen 7 3780U, 2.30 GHz, 4 cores; 16-GB RAM; Windows 10). The display had 2496 × 1664 pixels (201 ppi resolution), and the refresh rate was set to 60 Hz. The mouse was manufactured by Buffalo Inc., (model: BSMBU300, 1600 dpi). We used a standard-sized mousepad (21 cm × 17 cm).

The cursor speed was set as the default in the OS, i.e., the control-display gain was set at the middle of the slider in the Control Panel configuration. The pointer acceleration (or *Enhance pointer precision* setting in Windows 10) was enabled to allow the participants to perform mouse operations with higher ecological validity [10]. Fitts' law holds even when the pointer acceleration is turned on [1, 40].

4.2 Participants

Thirty unpaid students were collected from a local university (18 females and 12 males; ages: 18 to 26 years, $M = 22.2$, $SD = 1.59$). All belonged to a computer science department and were good at mouse operations. All had normal or corrected-to-normal vision. Twenty-six were right-handed and the remaining four were left-handed. PC usage history ranged from 0 (i.e., less than one year) to 19 years, $M = 7.63$, $SD = 5.31$.

4.3 Results

We recorded a total of $11_{PT} \times 4_W \times 16_{\text{repetitions}} \times 30_{\text{participants}} = 21{,}120$ data points. We applied the same criteria on outlier detection used in Experiment 1. There were no participant-level outliers and 70 trial-level outliers (0.33%); thus, we analyzed the remaining 21,050 data points.

Movement Time The Shapiro-Wilk test showed that 40 of the 44 conditions violated the normality assumption, so we log-transformed the data. We found significant main effects of PT ($F_{5.696,165.198} = 4.393$, $p < 0.001$, $\eta_p^2 = 0.13$) and W ($F_{1.725,50.024} = 910.0$, $p < 0.001$, $\eta_p^2 = 0.97$). The interaction of $PT \times W$ was not significant ($F_{12.982,376.475} = 1.310$, $p = 0.124$, $\eta_p^2 = 0.043$). Possibly due to the small sample size compared with the crowdsourcing study, there were only two pairs showing significant differences affected by PT (Figure 6a). A linear regression model of $MT = a + b \cdot ID_n + c \cdot PT$ yields adjusted $R^2 = 0.981$ and $c = 0.00458$ (with $p < 0.0001$).

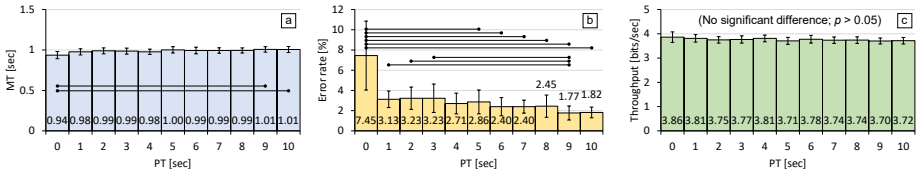

Fig. 6. Main effects of PT on (a) MT without log-transformation, (b) ER, and (c) throughput in Experiment 2. Error bars show 95% CIs, and horizontal lines indicate significantly different pairs (at least $p < 0.05$).

Table 3. Model fitness in Experiment 2. The yellow cell shows the lowest fit using ID_e. The pink cell indicates an insignificant ($p > 0.05$) contributor of ID_e to predict MTs. Blue cells indicate that there were significant differences in model fitness.

PT	Nominal				Effective				Difference		
[sec]	R^2	p value	AIC	BIC	R^2	p value	AIC	BIC	R^2	AIC	BIC
0	0.999	0.000619	-28.0	-29.2	0.159	0.601	-1.90	-3.13	0.840	-26.1	-26.1
1	0.987	0.00669	-18.0	-19.2	0.981	0.00943	-16.6	-17.8	0.00566	-1.38	-1.40
2	0.986	0.00682	-17.1	-18.4	0.984	0.00819	-16.4	-17.6	0.00240	-0.735	-0.763
3	0.981	0.00976	-15.9	-17.1	0.986	0.00714	-17.2	-18.4	-0.00543	1.28	1.26
4	0.996	0.00214	-22.0	-23.2	0.995	0.00272	-21.0	-22.3	0.000726	-0.991	-0.918
5	0.998	0.000799	-26.5	-27.7	0.995	0.00264	-21.7	-22.9	0.00340	-4.78	-4.81
6	0.986	0.00704	-17.1	-18.4	0.976	0.0122	-15.0	-16.2	0.00996	-2.15	-2.18
7	0.982	0.00917	-15.6	-16.9	0.978	0.0108	-15.0	-16.2	0.00374	-0.641	-0.669
8	0.991	0.00454	-18.6	-19.8	0.983	0.00878	-15.9	-17.2	0.00793	-2.68	-2.60
9	0.993	0.00353	-19.6	-20.8	0.989	0.00527	-18.0	-19.2	0.00395	-1.56	-1.58
10	0.997	0.00146	-23.4	-24.6	0.990	0.00503	-18.4	-19.7	0.00709	-4.99	-4.92

Error Rate We found significant main effects of PT ($F_{10,290} = 7.664$, $p < 0.001$, $\eta_p^2 = 0.20$) and W ($F_{3,87} = 9.939$, $p < 0.001$, $\eta_p^2 = 0.26$). The interaction of $PT \times W$ was significant ($F_{30,870} = 1.567$, $p < 0.05$, $\eta_p^2 = 0.051$). Similarly to Experiment 1, the ER for PT_0 was remarkably high, as shown in Figure 6b. There were no significant differences between any pair for $PT \geq 4$ sec. The mean ERs for W_{10} to W_{40} were 3.71, 2.82, 3.37, and 2.25%, respectively, and we found three pairs that showed significant differences ($p < 0.05$) for (W_{10}, W_{20}), (W_{10}, W_{40}), and (W_{30}, W_{40}).

Throughput and Fitts' Law Fitness We found a significant main effect of PT ($F_{10,290} = 1.949$, $p < 0.05$, $\eta_p^2 = 0.063$). There were no significant differences in any pair ($p > 0.05$, Figure 6c)[4]. The maximum difference was $3.86 - 3.70 = 0.16$ bits/sec (i.e., 4% of the baseline condition, PT_0). In comparison, in Experiment 1, the maximum difference was 0.11 bits/sec (3%), but there were several significantly different pairs (see Figure 3c). The results of this lab-

[4] It is possible for ANOVA that a main effect is significant but the pairwise tests show no significant differences.

Table 4. Results of Experiment 2. Blue cells indicate a violation of Fitts' law's assumption that W and SD_x are linearly related.

PT	MT [sec]				W_e [pixels]				R^2	ER	Throughput
[sec]	W_{10}	W_{20}	W_{30}	W_{40}	W_{10}	W_{20}	W_{30}	W_{40}	(W, SD_x)	[%]	[bits/sec]
0	1.13	0.954	0.869	0.794	74.8	76.2	67.1	93.7	0.300	5.24	3.68
1	1.20	0.981	0.887	0.850	23.7	45.2	69.6	90.4	0.999	2.25	3.67
2	1.23	0.986	0.904	0.842	24.6	46.7	70.2	87.7	0.997	1.96	3.67
3	1.22	0.985	0.883	0.852	23.9	48.0	68.8	88.6	0.998	1.80	3.64
4	1.21	0.991	0.904	0.829	23.7	46.5	69.4	89.8	0.999	2.05	3.63
5	1.21	1.01	0.922	0.850	24.2	45.7	68.6	92.8	0.999	1.70	3.61
6	1.22	0.981	0.916	0.828	23.7	44.7	68.9	86.9	0.997	1.76	3.61
7	1.24	0.982	0.901	0.840	24.2	46.4	69.9	89.1	0.998	2.00	3.58
8	1.23	1.02	0.883	0.846	24.1	46.2	66.7	91.7	0.998	1.65	3.64
9	1.26	1.02	0.925	0.854	23.0	45.4	68.1	91.6	1	1.40	3.57
10	1.24	1.02	0.914	0.851	23.3	44.1	67.0	91.2	0.999	1.61	3.58

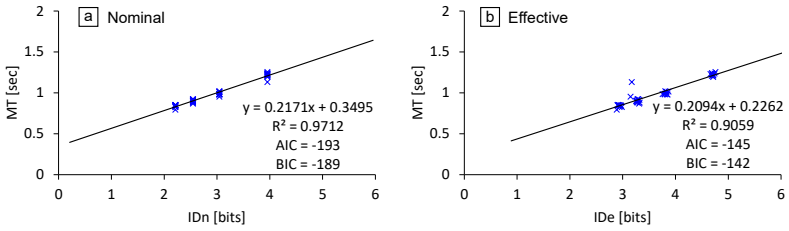

Fig. 7. Model fitness using (a) ID_n and (b) ID_e for 44 data points in Experiment 2.

based experiment again supported that throughput performance is not strongly affected by PT.

We found the same issues related to the effective width method as in Experiment 1. First, as shown in Table 3, the Fitts' law fitness using ID_e was remarkably low under the PT_0 condition ($R^2 = 0.159$, yellow cell), and ID_e was not a significant contributor ($p = 0.601$, pink cell). Second, the AIC and BIC differences between the ID_n and ID_e were significant only under the PT_0 condition (blue cells). Third, the correlation between W and SD_x was low only under the PT_0 condition ($R^2 = 0.300$), while the other PT conditions showed $R^2 > 0.99$ (Table 4). It seems that the W_es for W_{10} and W_{20} were high for PT_0 compared with the other PTs.

Overall, the results obtained in Experiment 1 were reproduced in this lab-based experiment. This result rejected our assumption that the remarkably low Fitts' law fitness using ID_e for PT_0 was due to the issues related to the crowdsourcing experiment such as instruction incompliance. It would be common that the endpoint distribution for small target widens than expected with the effective width method of Fitts' law if there is no PT. We again did not confirm the benefit to using ID_e for comparing different conditions as Figure 7 shows the lower model fitness for all 44 data points using ID_e.

5 Discussion

5.1 Effects of Time Penalty on Throughput, Model Fitness, and Endpoint Distributions

Our results indicate that crowdworkers and lab-based participants adjusted their strategy for pointing (speed and accuracy) in accordance with the given PT. The results indicate that, overall, a shorter PT is better for MT, although a too short PT negatively affects ER. The unified performance metric, throughput, was also higher for a shorter PT in Experiment 1 (Figure 3c), but there were no significant differences in Experiment 2 (Figure 6c). The throughputs fell within 3% in Experiment 1 and up to 4% in Experiment 2. Thus, the significant differences found in Experiment 1 were due to the larger sample size ($N = 125$ valid workers). We conclude that the normalization capability of throughput was partially demonstrated.

However, there is an issue related to the significantly low model fitness using ID_e under the zero-penalty condition. This was likely because the ER under the PT_0 condition was significantly higher, particularly for narrow targets, and the endpoint distribution SD_x (thus W_e) was widened. Because the effective width method is based on the assumption that SD_x widens accordingly to the nominal W, such a violation was harmful for Fitts' law fitness using ID_e.

Inconsistent with previous studies on the effective width method [37, 44], when we regressed all 44 data points ($11_{PT} \times 4_W$), the model fitness using ID_n was significantly better than ID_e in both experiments. This may be due to this lack of regularity between SD_x and W. Because there was a small effect of PT on MT (MT changed up to 7% from the baseline), using ID_n was sufficient for predicting MT regardless of PT. This is convenient for researchers and designers; when we arrange a new button/icon on a GUI, the only known parameter of the target is the nominal W; thus, we can use only the ID_n. This point has been mentioned in previous studies [37, 44], and we found that, for the penalty time paradigm, we can use ID_n for predicting MT even if there are several different PTs.

In comparison, if researchers would like to compare several input devices and user groups, we should use the accuracy-normalized model, but ID_e showed significantly worse fitness when regressing the 44 data points. This prevents us from recommending using ID_e, although it is recommended by the ISO standard [23]. One choice is to analyze the data separately under zero-penalty and non-zero penalty conditions. Because the zero-penalty condition has been used in previous studies and the ISO standard, comparing user groups for the zero-penalty data is not problematic. When we regressed the remaining 40 data points (PT_1 to PT_{10}) using ID_e, the R^2 for Experiments 1 and 2 were 0.979 and 0.938, respectively, which were worse than using ID_n but not considerably low. However, in realistic GUIs, there exist both zero-penalty empty space and time-consuming objects of mis-clicked like hyperlinks; thus, this separate analysis is not convenient for practitioners. More sophisticated methods need to be investigated for our future work.

5.2 Exception for Zero-penalty Condition

We found that the PT_0 condition resulted in remarkably different results, which is in line with the findings of Banovic et al. [3]. For example, in Banovic et al.'s MT prediction model, the prediction accuracy under all PT conditions showed $R^2 = 0.85$, but the data without PT_0 showed $R^2 = 0.98$ (see Figure 11 in [3]). Banovic et al.'s MT prediction model is based on the effective width method with which the endpoints follow a normal distribution and the SD_x increases accordingly to the given W. Thus, if this assumption is violated, the resultant MT prediction does not work well.

Because Banovic et al. did not report the endpoint distributions, we cannot discuss if this assumption is true for their data. However, the violation on SD_x was robustly found in both Experiments 1 and 2, and Banovic et al. reported that the ER for PT_0 was remarkably high (i.e., more endpoints fell outside the target). Therefore, we reasonably assume that this violation occurred and negatively affected Banovic et al.'s model fitness. Analyzing the resultant user behavior (here, SD_x) enables us to discuss the user strategy and corresponding high/low model fitness, which will lead to future model refinements.

5.3 Limitations and Future Work

Our findings are somewhat limited by the experimental conditions, e.g., we did not vary the target distance A and used a 1D target-pointing task. These conditions reflect our main focus, which was to observe the effects of PT on user performance measured as throughput. We chose this because throughput is independent of the choice of A and W [26, 35]. However, these limitations restrict the effective range of PT. For example, a single trial can be finished in 1 sec on average, but other GUI-operation tasks that take a longer time, such as navigation in hierarchical menus and webpage navigation, would likely result in different effects of PT on MT and ER.

Regarding the range of PTs (0 to 10 sec), real GUIs would have much longer PTs, as mentioned in Section 1, and thus the external validity is limited. This range was designed with reference to [3], which showed the effects of PT on MT and ER quickly plateau, and our experiments were internally valid. The ecological validity is also limited. For example, when an unintended hyperlink is clicked, we assume a recovery time is needed. In real GUIs, however, typically there is a margin between links; clicking this area induces no PT. Still, our contribution is bridging the gap between tested conditions and real GUIs, which was inadequate in previous studies using only PT_0 conditions.

It is known that devices [9], handedness [19], and ages [18] affect pointing performance. We assumed that evaluating these factors did not strengthen our novelty and contribution, and thus we did not analyze them. However, our tested platform utilizes comparatively older workers than (e.g.) Amazon MTurk: the former has 20% of <30-year-old people, while the latter has 30% [30, 34]. Testing the generalizability of our findings to other countries' platforms will be included in our future work.

In particular for the $PT = 0$ sec condition, clicking repeatedly around the target could result in finishing the task more quickly. For crowdworkers, such a strategy is effective to maximize payment per time, which might affect the results. Also, we mainly discussed the quantitative data, but a qualitative analysis based on interviews related to subjective strategies would uncover the reasons behind the differences in results.

We found that the model fitness using ID_n and ID_e for non-zero penalties was not significantly different in accordance with AIC and BIC. Also, when we analyzed the 44 data points in a mixed manner, the fit for ID_n was better than ID_e. These findings are inconsistent with previous studies on the effective width method [37, 44], but this might be because we used only one A and a limited number of Ws. While it is assumed with the effective width method that the SD_x is not affected by A [26, 35], it is slightly affected by the nominal A [42, 44]. Therefore, we cannot conclude that using ID_e shows comparable model fitness with ID_n if there are several target distances. This informs possible future work that examining if the same findings can be observed.

6 Conclusion

We explored the effects of penalty time PT on user performance in terms of the throughput of target-pointing tasks. The throughputs decreased as PT increased in the crowdsourcing experiment, while it was not significantly affected in the lab-based experiment. Because the throughput difference was up to 3 and 4% in these experiments, respectively, we partially found the normalization capability of the effective width method of Fitts' law [23, 28, 35]. However, we also found that using a 0-sec PT resulted in different outcomes under the other conditions, particularly when a high error rate was observed and Fitts' law fitness using ID_e was low, which makes estimating throughput performance difficult. This would make it difficult to model the movement time and error rate as Banovic et al. found [3], because participants behave differently only under the zero-penalty condition. This difficulty probably comes from the fact that the endpoint variability SD_x violates the assumed linear relationship to the nominal W, which was consistently found in the both experiments.

Our results filled the gap between the standardized methodology of pointing performance [23, 35] and the penalty-time paradigm for realistic GUIs [3, 39]. Particularly, the data obtained in our experiments contributed to our better understanding of user behavior in GUIs that would induce time penalty. Thus, this work is a necessary step towards refining current performance prediction models, which may be helpful to researchers and designers in the future. Last, if we look at only the PT_0 condition, using ID_n yields $R^2 = 0.996$ and 0.999 in the two experiments, but using ID_e yields significantly lower fits; this is consistent with previous studies. Comparing PT_0 with the other PT_s highlighted the unsuitability of using PT_0, but we note that there were also consistencies with previous studies.

References

1. 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
2. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6), 716–723 (Dec 1974). https://doi.org/10.1109/TAC.1974.1100705
3. Banovic, N., Grossman, T., Fitzmaurice, G.: The effect of time-based cost of error in target-directed pointing tasks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1373–1382. CHI '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2470654.2466181, http://doi.acm.org/10.1145/2470654.2466181
4. Berinsky, A.J., Huber, G.A., Lenz, G.S.: Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. Political Analysis **20**(3), 351–368 (2012). https://doi.org/10.1093/pan/mpr057
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. Bi, X., Zhai, S.: Bayesian touch: a statistical criterion of target selection with finger touch. In: Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '13). pp. 51–60 (2013). https://doi.org/10.1145/2501988.2502058
7. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science **6**(1), 3–5 (2011). https://doi.org/10.1177/1745691610393980
8. Burnham, K.P., Anderson, D.R.: Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media (2003)
9. Card, S.K., English, W.K., Burr, B.J.: Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a crt. Ergonomics **21**(8), 601–613 (1978). https://doi.org/10.1080/00140137808931762
10. Casiez, G., Roussel, N.: 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. pp. 603–614. UIST '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2047196.2047276, http://doi.acm.org/10.1145/2047196.2047276
11. Crossman, E.R.: The speed and accuracy of simple hand movements. Ph.D. thesis, University of Birmingham (1956)
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. Fitts, P.M., Radford, B.K.: Information capacity of discrete motor responses under different cognitive sets. Journal of experimental psychology **71**(4), 475–482 (April 1966). https://doi.org/10.1037/h0022970
15. Gillan, D.J., Bias, R.G.: Fitting motivation to fitts' law : Effect of a penalty contingency on controlled movement. Proceedings of the Human Factors and Ergonomics Society Annual Meeting **62**(1), 265–269 (2018). https://doi.org/10.1177/1541931218621061
16. Gori, J., Rioul, O., Guiard, Y.: Speed-accuracy tradeoff: A formal information-theoretic transmission scheme (fitts). ACM Trans. Comput.-Hum. Interact. **25**(5), 27:1–27:33 (Sep 2018). https://doi.org/10.1145/3231595, http://doi.acm.org/10.1145/3231595
17. Gori, J., Rioul, O., Guiard, Y., Beaudouin-Lafon, M.: The perils of confounding factors: How fitts' law experiments can lead to false conclusions. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3173574.3173770, https://doi.org/10.1145/3173574.3173770
18. Hertzum, M., Hornbæk, K.: How age affects pointing with mouse and touchpad: A comparison of young, adult, and elderly users. International Journal of Human-Computer Interaction **26**(7), 703–734 (2010). https://doi.org/10.1080/10447318.2010.487198
19. Hoffmann, E.R.: Movement time of right- and left-handers using their preferred and non-preferred hands. International Journal of Industrial Ergonomics **19**(1), 49–57 (1997). https://doi.org/10.1016/0169-8141(95)00092-5
20. 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
21. Horton, J.J., Rand, D.G., Zeckhauser, R.J.: The online laboratory: conducting experiments in a real labor market. Experimental Economics **14**(3), 399–425 (Sep 2011). https://doi.org/10.1007/s10683-011-9273-9, https://doi.org/10.1007/s10683-011-9273-9
22. Huang, J., Tian, F., Fan, X., Zhang, X.L., Zhai, S.: Understanding the uncertainty in 1d unidirectional moving target selection. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3173574.3173811, https://doi.org/10.1145/3173574.3173811
23. ISO: Iso 9241-9. international standard: ergonomic requirements for office work with visual display terminals (vdts)–part 9: requirements for non-keyboard input devices, international organization for standardization (2000)
24. Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association **90**(430), 773–795 (1995). https://doi.org/10.1080/01621459.1995.10476572
25. 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

26. 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
27. 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, Association for Computing Machinery, New York, NY, USA (1992). https://doi.org/10.1145/142750.142794, https://doi.org/10.1145/142750.142794
28. 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
29. MacKenzie, I.S., Sellen, A., Buxton, W.A.S.: A comparison of input devices in element pointing and dragging tasks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 161–166. CHI '91, Association for Computing Machinery, New York, NY, USA (1991). https://doi.org/10.1145/108844.108868, https://doi.org/10.1145/108844.108868
30. Moss, A.: Demographics of people on amazon mechanical turk (2020), retrieved April 4, 2021 from https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/
31. Olafsdottir, H.B., Guiard, Y., Rioul, O., Perrault, S.T.: A new test of throughput invariance in fitts' law: Role of the intercept and of jensen's inequality. In: Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers. p. 119–126 (2012)
32. Ren, X., Zhou, X.: An investigation of the usability of the stylus pen for various age groups on personal digital assistants. Behaviour & Information Technology **30**(6), 709–726 (2011). https://doi.org/10.1080/01449290903205437
33. 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
34. Shimizu, N., Nakagawa, M.: Crowdsourcing: Current status and potential: 2. current trends and issues in microtask-based crowdsourcing. IPSJ Magazine **56**(9), 886–890 (aug 2015)
35. 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
36. 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
37. Wright, C.E., Lee, F.: Issues related to hci application of fitts's law. Human-Computer Interaction **28**(6), 548–578 (2013). https://doi.org/10.1080/07370024.2013.803873

38. Yamanaka, S.: Effect of gaps with penal distractors imposing time penalty in touch-pointing tasks. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services. MobileHCI '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3229434.3229435, https://doi.org/10.1145/3229434.3229435
39. Yamanaka, S.: Risk effects of surrounding distractors imposing time penalty in touch-pointing tasks. In: Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces. pp. 129–135. ISS '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3279778.3279781, http://doi.acm.org/10.1145/3279778.3279781
40. Yamanaka, S.: Evaluating temporal delays and spatial gaps in overshoot-avoiding mouse-pointing operations. In: Proceedings of Graphics Interface 2020. pp. 440 – 451. GI 2020 (2020). https://doi.org/10.20380/GI2020.44
41. 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
42. Yamanaka, S., Usuba, H.: Rethinking the dual gaussian distribution model for predicting touch accuracy in on-screen-start pointing tasks. Proc. ACM Hum.-Comput. Interact. **4**(ISS) (Nov 2020). https://doi.org/10.1145/3427333, https://doi.org/10.1145/3427333
43. Zhai, S.: Characterizing computer input with fitts' law parameters – the information and non-information aspects of pointing. International Journal of Human-Computer Studies **61**(6), 791–809 (2004). https://doi.org/10.1016/j.ijhcs.2004.09.006, fitts' law 50 years later: applications and contributions from human-computer interaction
44. 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