Investigation of Test-Retest Reliability in 2D-Target Pointing Task for Five Consecutive Days

YUI ATARASHI, University of Tsukuba, Japan
YUKI TAKEYAMA, University of Tsukuba, Japan
MAYU AKATA, University of Tsukuba, Japan
YUMA SUZUKI, University of Tsukuba, Japan
SHINGO KATO, University of Tsukuba, Japan
KAISEI YOKOYAMA, University of Tsukuba, Japan
TAKUMA HIDAKA, University of Tsukuba, Japan
SHOTA YAMANAKA, LY Corporation, Japan
BUNTAROU SHIZUKI, University of Tsukuba, Japan

Test-retest reliability means whether the same results can be observed when the same participants perform the same task over two sessions. By considering test-retest reliability, researchers can review the reliability of their study and its results. Previous studies showed that 1D-target pointing tasks have low test-retest reliability. However, it should be more practical to investigate test-retest reliability for 2D-target pointing tasks because the situation represents a more realistic use case than for 1D-target pointing tasks. We asked the participants to perform a 2D-target pointing task for five days under identical conditions and then to answer a detailed questionnaire. The result showed that the researchers should assemble a study with three or more sessions to make the test-retest reliability high in the aspect of movement time, error rate, and throughput.

CCS Concepts: • **Human-centered computing** → **HCI theory, concepts and models**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Fitts' law, pointing, graphical user interface, human motor performance

ACM Reference Format:
Yui Atarashi, Yuki Takeyama, Mayu Akata, Yuma Suzuki, Shingo Kato, Kaisei Yokoyama, Takuma Hidaka, Shota Yamanaka, and Buntarou Shizuki. 2024. Investigation of Test-Retest Reliability in 2D-Target Pointing Task for Five Consecutive Days. 1, 1 (November 2024), 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' addresses: Yui Atarashi, University of Tsukuba, Tokyo, Japan, atarashi@iplab.cs.tsukuba.ac.jp; Yuki Takeyama, University of Tsukuba, Ibaraki, Japan, takeyama@iplab.cs.tsukuba.ac.jp; Mayu Akata, University of Tsukuba, Ibaraki, Japan, akata@iplab.cs.tsukuba.ac.jp; Yuma Suzuki, University of Tsukuba, Ibaraki, Japan, yusuzuki@iplab.cs.tsukuba.ac.jp; Shingo Kato, University of Tsukuba, Ibaraki, Japan, skato@iplab.cs.tsukuba.ac.jp; Kaisei Yokoyama, University of Tsukuba, Ibaraki, Japan, kyokoyama@iplab.cs.tsukuba.ac.jp; Takuma Hidaka, University of Tsukuba, Ibaraki, Japan, hidaka@iplab.cs.tsukuba.ac.jp; Shota Yamanaka, LY Corporation, Tokyo, Japan, syamanak@yahoo-corp.jp; Buntarou Shizuki, University of Tsukuba, Ibaraki, Japan, shizuki@cs.tsukuba.ac.jp.

1 INTRODUCTION

1.1 Background

Test-retest reliability means whether we can observe the same results when the same participants perform the same task over two sessions [37]. Researchers often form their conclusions after analyzing the results of one-time user experiments. However, in such cases, good (or bad) results obtained by chance would lead to wrong conclusions. By considering test-retest reliability, researchers can more accurately review the reliability of the experimental results. Test-retest reliability has been the focus of attention since around 1937, particularly in the psychology field [12, 13, 28, 33]. For example, previous researchers [28] have investigated whether the same participant group provides the same responses when they are given the same questionnaire after an interval of several weeks.

In the HCI field, many studies emphasize the importance of replication (e.g., [5, 17, 34]). On the other hand, few studies have investigated test-retest reliability. In replication studies, researchers focus on experimenting under the same experimental conditions (e.g., equipment, participant demographics, and tasks). In contrast, test-retest reliability expands the focus to the same participants in addition to the same experimental conditions. Ideally, the same participants should give the same results over two sessions. However, in the psychology field, several researchers have reported cases in which they did not obtain the same results [12, 13, 28, 33].

Although the exploration of test-retest reliability in the HCI field remains limited, initial attempts have emerged. These studies have primarily focused on simpler tasks, such as 1D-target pointing, as a starting point. Sharif et al. [29] and Yamanaka [37] investigated test-retest reliability in 1D-target pointing tasks and found that it was low. However, we believe it is beneficial to investigate test-retest reliability for 2D-target pointing tasks because it represents a more realistic use case than 1D-target pointing tasks. For the 2D-target pointing task, previous researchers [19] investigated a multi-session experiment, although they did not investigate the test-retest reliability of the task. This lack of clarity regarding the test-retest reliability in a realistic task condition (2D-target pointing) is what motivated us to conduct this study.

Since a target-pointing task requires participants to perform rapid aimed movements, we suspect that the test-retest reliability is affected by the participants' factors, i.e., their physical condition on the test day or their improved skill in mouse manipulation. For this reason, we asked participants to perform the multi-directional pointing task shown in the ISO9241–411 [11] for five days under identical conditions and then to answer a detailed questionnaire. We then analyzed the results to determine which factors reduced the test-retest reliability. We also examined whether it is necessary to repeat the same task for more than two sessions in order to improve the test-retest reliability.

1.2 Research Questions

We address the following two research questions (RQs) in this research.

RQ1 How many sessions should the researchers conduct to stabilize the participant's performance on the 2D-target pointing task?

RQ2 Which factors have an effect on participant performance?

1.3 Contribution Statement

The contributions of this study are twofold.

- We evaluated test-retest reliability in five sessions conducted for five consecutive days in an offline environment with fully consistent equipment.

- We investigated which questionnaire items are necessary for evaluating test-retest reliability in pointing tasks.

2 RELATED WORK

In this section, we first describe Fitts' law and throughput [10, 21, 31, 36], which we use to evaluate test-retest reliability in this study. We then present studies on test-retest reliability in a 1D-target pointing task. Finally, we show studies comparing the performance between two sessions, and how we evaluate the performance of all the sessions in this study.

2.1 Fitts' Law and Throughput

Fitts' law models the time required to select a target using an input device such as a mouse [10]. In this paper, we use Fitts' law defined as Equation (1) [21]:

$$MT = a + b \times ID, \text{ with } ID = \log_2\left(\frac{D}{W} + 1\right), \tag{1}$$

where MT is the movement time, D is the distance to the target, W is the width of the target, and a and b are empirical constants. ID (index of difficulty) is the difficulty of target selection.

When comparing multiple input devices, pointing methods, or user groups, the definition of the high-performance device/method/group might change depending on the MT and ER (error rate). For example, if there are two devices with different MT and ER, the definition of a high-performance device changes depending on the importance of MT or ER. Alternatively, if the ER of the two devices has the same value, we can define a superior device by comparing only the MT. Using TP (throughput) is recommended when performing such comparisons [31], as it helps resolve this issue. This metric integrates speed and accuracy and is calculated using ID_e (effective index of difficulty) consisting of W_e (effective width) that considers click-point variability [6], which is defined as

$$ID_e = \log_2\left(\frac{D}{W_e} + 1\right) \text{ and } W_e = \sqrt{2\pi e}\sigma = 4.133\sigma, \tag{2}$$

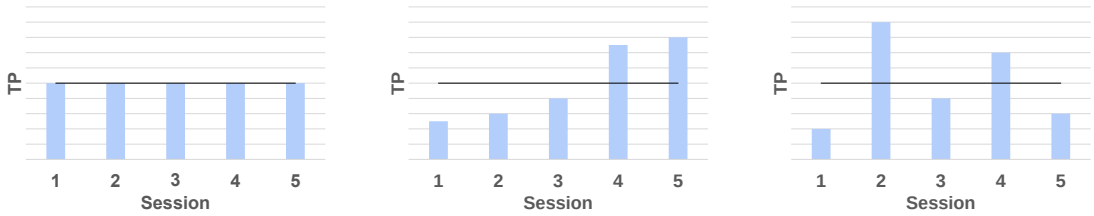
where σ is the standard deviation of the click-point distribution. W_e is adjusted so that ER becomes 4%. A common definition of TP as shown in Equation (3) is called the mean-of-means method [31, 36], which has also been used in previous studies on the test-retest reliability of pointing tasks [29, 37]:

$$TP = \frac{1}{|A| \times |W|} \sum_{i=1}^{|A|\times|W|} \left(\frac{ID_{e_i}}{MT_i}\right), \tag{3}$$

where i represents the i-th task condition within $|A| \times |W|$. Calculating TP using W_e can reduce the effect of participants' subjective biases toward speed and accuracy. While several previous studies used the bivariate SD of click coordinates on a 2D plane as σ [3, 38], utilizing a univariate SD along the task axis is theoretically and empirically valid [30, 36]. In this study, we use a univariate SD to compute TP. The robustness of TP is the reason that the ISO9241–411 [11] recommends its use, and why numerous researchers calculate TP when investigating the performance of multiple devices, methods, and groups.

2.2 Test-retest Reliability in 1D-Target Pointing Task

As mentioned earlier, test-retest reliability has been investigated extensively in the field of psychology (e.g., [12, 13, 28, 33]), but it has received much less attention in the target-pointing task in the HCI field. Two of the few related



(a) A pattern where *SV* is zero. (b) A pattern where *SV* is low. (c) A pattern where *SV* is high.

Fig. 1. Correspondence between the data collected from each session and the *SV*. As shown by the black line, the three averages are equal. However, the more non-linear the variance in performance, the larger the *SV*.

works were done by Sharif et al. [29] and Yamanaka [37]. Sharif et al. [29] investigated the test-retest reliability in 1D-target pointing tasks in terms of *TP*. They found that the test-retest reliability was low because some participants' *TP* differed between sessions while others remained stable, and the mean of *TP* for all participants significantly differed between sessions. Yamanaka [37] investigated test-retest reliability in terms of *MT*, *ER*, and *TP*, and found that the test-retest reliability was low because the *MT*, *ER*, and *TP* of some participants significantly differed between sessions. Furthermore, Yamanaka [**?**] explored a method in a comparative experiment based on user interfaces by introducing a group sequential design, in which the p-value is checked at the midpoint of the experiment to determine whether to end the experiment early or to proceed.

Considering actual use cases, most pointing tasks are two-dimensional. Therefore, we investigate test-retest reliability for 2D-target pointing tasks based on the ISO9241−411 [11]. Moreover, while both studies mentioned above compared the performances of two sessions, a larger number of sessions would provide further insight into the changes in participants' performance. We, therefore, conduct our experiment for five sessions. In addition, Yamanaka's study used crowdsourcing for the experiment, but it was unclear whether the participants appropriately followed the instructions (e.g., to use a single input device throughout the experiment). It has also been reported that the accuracy of a task is lower when experiments are conducted remotely compared to when they are conducted on-site [9]. For these reasons, we investigate the test-retest reliability by conducting the experiments on-site.

Another difference between our experiment and these previous studies is the control of interval length. Sharif et al. [29] conducted two sessions with intervals ranging from four to 48 hours. The actual interval was arbitrarily chosen by each participant and thus not well-controlled. Yamanaka [37] compared two participant groups, one with intervals of one to two minutes and the other with intervals of 20 to 30 minutes. These intervals were relatively short, and he mentioned it as a limitation. In comparison, we controlled the start time of each session more strictly. Specifically, we asked the participants to initiate the task at an almost fixed time every day, and the result of the difference in start times was less than 15 minutes.

2.3 Metrics to Compare Performance across Sessions

Similar to previous studies [29, 37], we compare participants' performance between sessions in terms of *MT*, *ER*, and *TP*. The stability of the model fits using Fitts' law is not the subject of our current study; therefore, the results (regression expressions and R^2 values) are included in the supplementary materials.

To quantitatively evaluate the test-retest reliability, we use Matejka et al.'s method [24]. They investigated the effect of the appearance of a visual analog scale on the rating behavior of survey respondents. The smoothness of the survey results was evaluated using Equation (4), which is a variant of standard deviation. By applying this equation to the five sessions of data collected from the participants in this study, we can evaluate the performance improvement over the entire session. For example, the bias for participants whose performance improves linearly with the number of sessions becomes close to zero. On the other hand, the bias for participants whose performance improves or worsens non-linearly with the number of sessions becomes large; see Fig. 1. Considering the meaning of Equation (4) in this study, we refer to it as session variance (*SV*).

$$SV = \sqrt{\frac{1}{4}\sum_{i=1}^{4}\left((x_{i+1} - x_i) - \frac{\sum_{i=1}^{4}(x_{i+1} - x_i)}{4}\right)} \tag{4}$$

The metrics such as the daily mean of *MT*, *ER*, and *TP* can be applied to x.

3 EXPERIMENT

We conducted a 2D-target pointing experiment consisting of five sessions with 17 participants. Compared to previous studies in terms of session count (e.g., two sessions used by Sharif et al. [29] and Yamanaka [37]) or duration (e.g., the two-day experiment by Card et al. [4]), we aimed to include as many sessions as possible. However, we decided that weekend participation would pose a considerable burden on participants, and therefore we chose to conduct one session each day over five weekdays.

3.1 Apparatus

We used a laptop PC (Intel Core i7-11800H, 16 GB RAM, Windows 11, NVIDIA GeForce RTX 3060 LapTop GPU). The screen size was 15.6 inches, with dimensions of 359.7 × 227.4 mm, a resolution of 1920 × 1080 pixels, 0.179 mm per pixel, and a refresh rate of 60 Hz. The mouse was Microsoft Mobile Mouse 3500 (Wireless, 1000 DPI, 1000 Hz polling rate). The experimental system was developed with Compose for Desktop[1] and Kotlin[2], and displayed in full-screen mode.

3.2 Participants

A total of 17 individuals participated in the experiment (18–24 years old, M = 22.3, SD = 1.67). Thirteen were male, and four were female. One participant was left-handed, 16 were right-handed, and none were ambidextrous.

We determined the sample size (i.e., the number of participants) using G*Power [8]. As a primary performance metric is *TP* and a previous study reported that the effect size of the session on *TP* was somewhere between *medium* and *large* depending on the interval length [37], we set Cohen's $f = 0.325$, power = 0.8, and $\alpha = 0.05$ for RM-ANOVAs. We found that a minimum of 13 participants were required. To allow for the possibility that several participants might drop out during the five-day experiment, we decided to recruit more than 13 participants.

3.3 Task, Design, and Procedure

To measure pointing performance trends with a high degree of accuracy, it is recommended that 15–25 trials be performed for each target condition [31]. Thus, 25 circular targets were displayed (Fig. 2) in the task window (1920 × 1080 pixels). The current active target was colored in red and the inactive ones in gray. If participants clicked on the active target,

[1]https://www.jetbrains.com/lp/compose-desktop/
[2]https://kotlinlang.org/

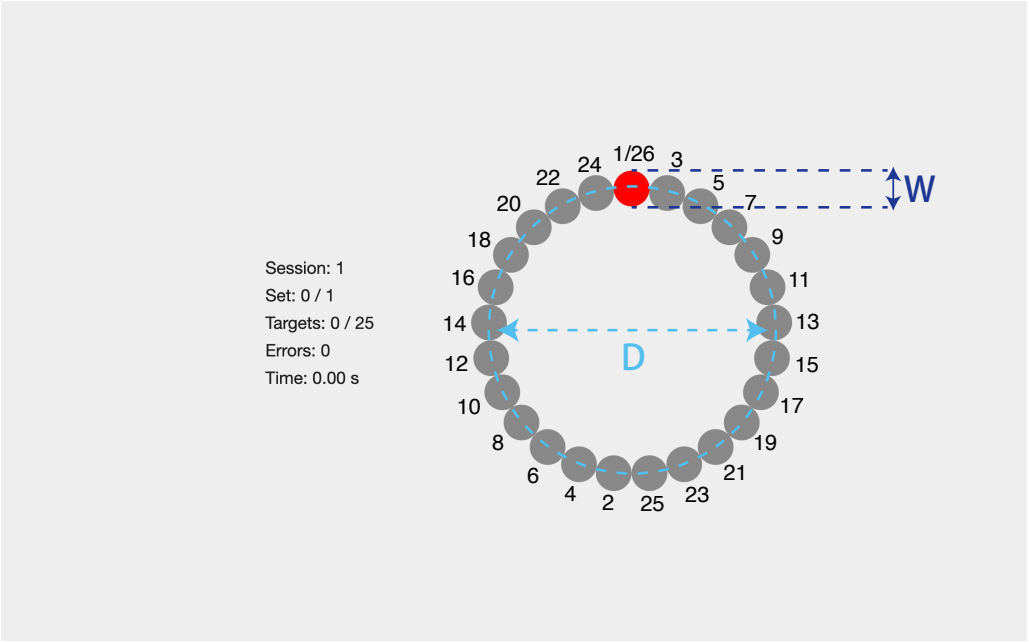


Fig. 2. Appearance of the application used in the experiment and the order in which targets are clicked.

the next target turned red. If participants clicked outside the active target, it flashed orange, and they had to keep trying until successfully clicked it. One set consisted of 25 targets for a fixed $D \times W$ condition. The first target was on the top.

The independent variables were target distance D (300, 440, and 630 pixels), target width W (8, 20, 38, and 78 pixels), and Session (1 to 5). The D and W conditions were selected concerning Yamanaka et al. [38] so that ID ranged from 2.28 to 6.32 bits. The dependent variables were movement time (MT), error rate (ER), and throughput (TP).

The order of the 12 $D \times W$ conditions was fixed in ascending order of ID to control the effect of order among participants. Before the first set of each session, participants got used to the task by performing a practice set with D = 400 and W = 30 pixels, the results of which were not used for analysis.

Before the practice set of each session, participants completed the pre-session questionnaire consisting of seven items and completed the post-session questionnaire consisting of one item after each session (Table 1). There are numerous factors that could affect pointing performance between sessions, and thus it is desirable to ask for as much information as possible. Note that since factors affecting test-retest reliability in pointing tasks have never been established, it is difficult to determine concrete factors in advance. Therefore, we can only know whether these factors are effective or not after we analyze them. It is thus meaningful to come up with as many questionnaire items as possible. For example, we wanted to ask questions in advance of the first session about how long the participants usually use a Windows PC, their usual mouse cursor-speed settings, their usual display sizes (resolutions and inches), their usual display brightness, and their usual chair heights (if the chair used in the experiment was extremely lower, the performance might be degraded). In addition, before every session, we wanted to ask how long the participants had exercised on the day before the experiment, how long they operated the mouse on a different PC, and how long and at what load they used a smartphone. However, it was not realistic to ask all possible questions every session due to time constraints, so we

included only limited items in the questionnaire, which was designed to find out what is worth investigating more deeply in the future. Questionnaire items for sleep, hunger, fatigue, and busyness were designed with reference to the paper [25] on health status and performance.

After finishing the five sessions, participants completed a post-experiment questionnaire (Table 2), asking participants' attributes such as their age, sex (free-form to allow arbitrary answers), dominant hand, and history of mouse use. The items related to games were set up with reference to the research result [27] that gamers have higher performance than non-gamers. The questionnaire items were also set up with reference to [32] for age, [2] for sex, and [20] for dominant hand and time spent using the mouse.

3.4 Interval Length

The participants completed one session per day five days, completing a total of five sessions. Before participating in the experiment, each participant chose his/her own start time (e.g., 1 p.m.) for five sessions to control the interval length. The largest difference in actual session start times was 14 minutes and 11 seconds. Thus, there were approximately 24 hours between the sessions for each participant. This allowed participants to recover from any fatigue caused by the previous session before the next session started.

4 RESULTS

MT was the duration from when the previous target was successfully clicked to when the next click was performed [18, 31]. Trials in which we observed one or more clicks outside the active target were flagged as errors.

Our *MT* and *ER* data did not pass the Shapiro-Wilk normality test ($\alpha = 0.05$). Therefore, we conducted a non-parametric ANOVA with the ART (Aligned Rank Transform) [16, 26, 35] for *MT* and *ER*. Additionally, to further explore the effects, we applied ART-C (Aligned Rank Transform for Contrasts) [7] for specific contrasts within *MT* and *ER*. For *TP*, as *TP* merged *D* and *W*, we considered only sessions as the independent variable. The Shapiro-Wilk normality test indicated that the *TP* data followed a normal distribution, and thus we used RM-ANOVAs with the Bonferroni p-value adjustment method for pair-wise tests.

4.1 Outlier Data Screening

We removed spatial outliers if the distance of the first click position was shorter than $D/2$ [1, 23] to remove clear accidental operations such as double-clicking the previous target. We also removed those that were farther than $2W$ from the target center.

Among the 25500 trials ($= 3_D \times 4_W \times 25_{clicks} \times 5_{sessions} \times 17_{participants}$), we removed 46 trial-level outliers (0.180%). While Sharif et al. removed data from four participants whose *ER*s were greater than 8% [29], which is twice the common *ER* [31], we did not remove any participants based on *ER*. This was because, although it has been argued that the *ER* in pointing tasks is 4% [22, 31], this is arbitrary, and the *ER* might actually be affected by task conditions such as target size [14].

4.2 MT

The *MT* results are shown in Fig. 3 and 4. *D*, *W*, and sessions are independent variables for the statistical analysis. We found significant main effects of *D* ($F_{2,32} = 593, p < 0.001, \eta_p^2 = 0.971$), *W* ($F_{3,48} = 1569.07, p < 0.001, \eta_p^2 = 0.990$), and sessions ($F_{4,64} = 18.71, p < 0.001, \eta_p^2 = 0.540$). Posthoc tests showed significant differences between session 1–3 ($p < 0.001$), 1–4 ($p < 0.01$), 1–5 ($p < 0.01$), 2–3 ($p < 0.001$), 2–4 ($p < 0.01$), and 2–5 ($p < 0.01$). The significant interaction

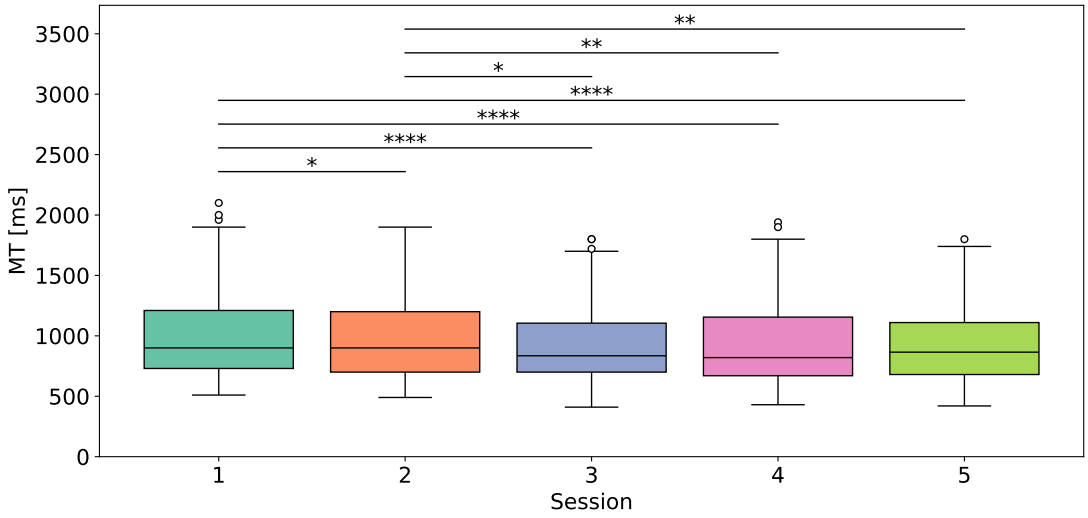


Fig. 3. *MT* of each session (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$).

($p < 0.05$) was found for $D \times W$ ($F_{6,96} = 2.06, p < 0.021, \eta_p^2 = 0.150$) and $W \times$ sessions ($F_{12,192} = 18.71, p < 0.001, \eta_p^2 = 0.114$).

4.3 ER

The *ER* results are shown in Fig. 5 and 6. D, W, and sessions are independent variables for the statistical analysis. We found a significant main effect of D ($F_{2,32} = 4.24, p = 0.023, \eta_p^2 = 0.210$), W ($F_{3,48} = 85.33, p < 0.001, \eta_p^2 = 0.842$) and Session ($F_{4,64} = 61.9, p = 0.004, \eta_p^2 = 0.209$). Significant interaction was shown between $W \times$ Session ($F_{12,192} = 3.21, p = 0.0003, \eta_p^2 = 0.167$).

4.4 TP

The *TP* results are shown in Fig. 7 and 8. D and W are taken into consideration to *TP*; thus, only sessions are the independent variable for the statistical analysis.

We found a significant main effect of sessions ($F_{4,64} = 33.143, p < 0.001, \eta_p^2 = 0.684$). Bonferroni-corrected pairwise comparisons revealed statistically significant differences between session 1–2 ($p < 0.01$), 1–3 ($p < 0.001$), 1–4 ($p < 0.001$), 1–5 ($p < 0.001$), 2–3 ($p < 0.001$), 2–4 ($p < 0.001$), and 2–5 ($p < 0.01$). No significant differences were found between sessions 3–4, 3–5, and 4–5.

4.5 Correlation between Participants' Performance and Questionnaire Results.

Table 1 shows the results of the correlation between participants' performance in each session and the answers to the session questionnaire. The correlation values were interpreted as follows [15]: 0.0–0.2 little if any; 0.2–0.4 weak; 0.4–0.7 moderate; 0.7–1.0 strong. The key finding are as follows. *MT* was weakly correlated with the current sleepiness. *MT*, *ER*, and *TP* were weakly correlated with current hunger level. *MT* and *TP* were weakly correlated with fatigue of the dominant hand.

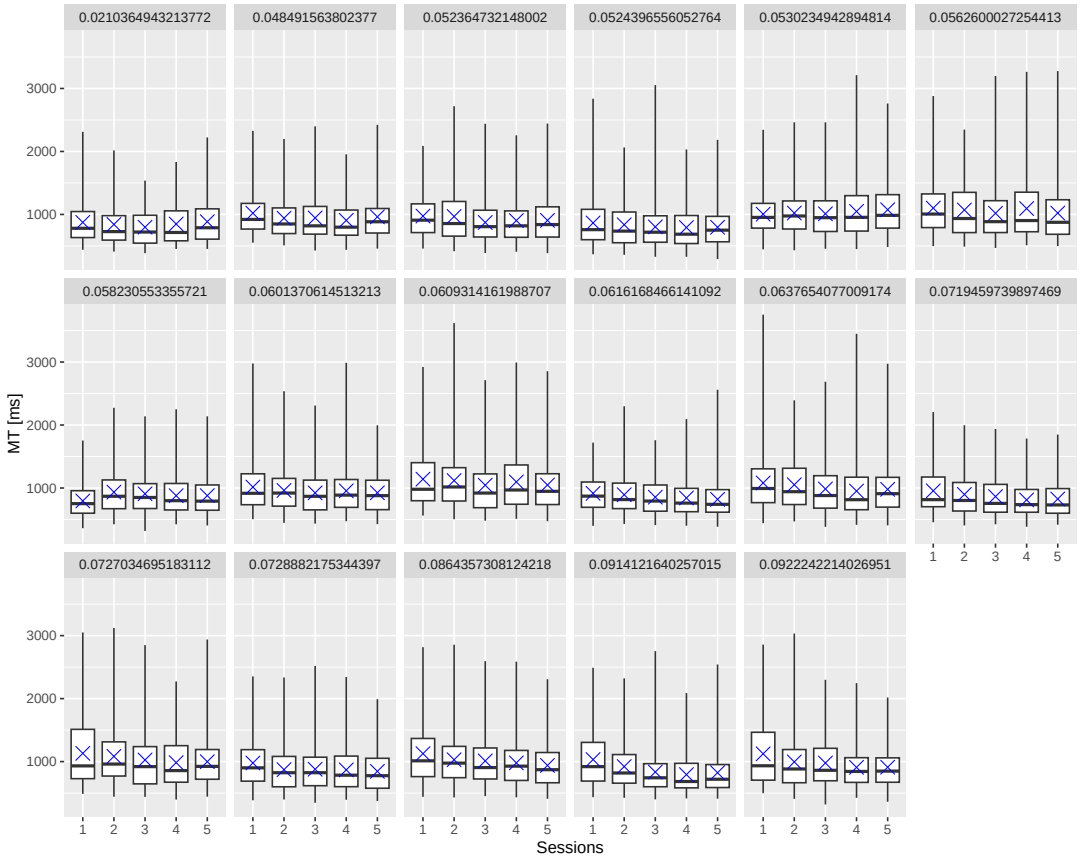


Fig. 4. *MT* of each session and participant. The title of each graph means *SV*.

Table 2 shows the results of the correlation between the average of participants' performance in each session and answers to the post-experiment questionnaire. The key findings are as follows. *SV* (*ER*) was weakly correlated and *SV* (*MT*) and *SV* (*TP*) were moderately correlated with whether or not participants played games using a mouse. *MT*, *TP*, *SV*(*ER*), and *SV*(*TP*) were weakly correlated and *SV*(*MT*) was moderately correlated with how long the mouse was used per day.

5 DISCUSSION

5.1 RQ1. How Many Sessions Should the Researchers Conduct to Stabilize the Participant's Performance on the 2D-Target Pointing Task?

The results indicate that the answer is *three* sessions. According to Fig. 3, there were significant differences in *MT* between Session 1 and Sessions 2, 3, 4, and 5, as well as between Session 2 and Sessions 3, 4, and 5. Therefore, *MT* is stabilized from *three* sessions. According to Fig. 5, *ER* in the first session showed statistically significant differences with the Sessions 2, 4, and 5. Thus, *ER* is stabilized from *two* sessions. Additionally, Fig. 3 shows that *TP* exhibited significant differences between several sessions. Specifically, significant differences were observed between Session 1 and Sessions

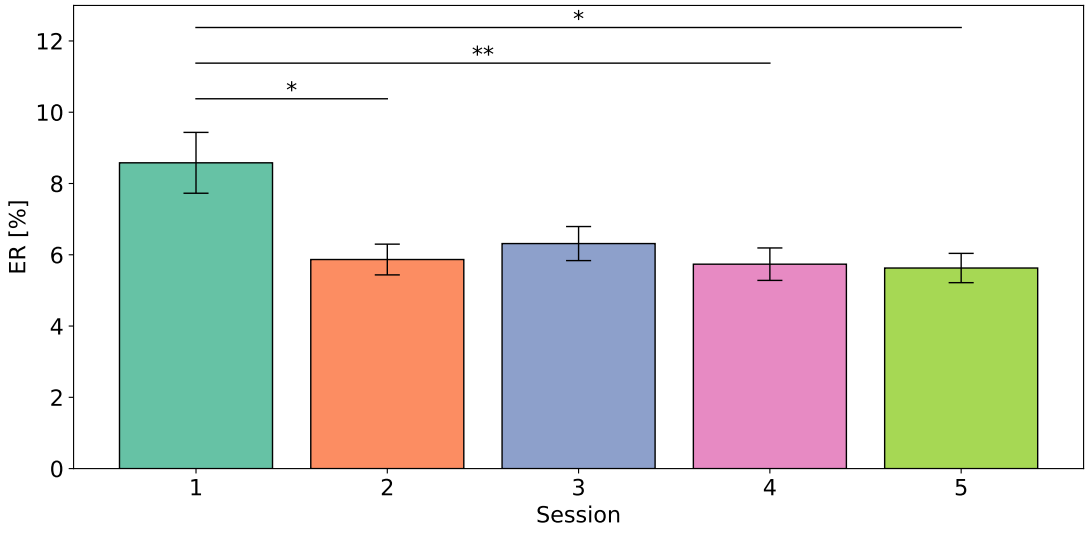


Fig. 5. *ER* of each session (Error bars represent Standard Error, SE; *: $p < 0.05$, **: $p < 0.01$).

2, 3, 4, and 5, as well as between Session 2 and Sessions 3, 4, and 5. However, no significant differences were found between Sessions 3, 4, and 5. This indicates that *TP* stabilized after *two* sessions.

These results suggest that three or more sessions should be conducted to stabilize *MT*, *ER*, and *TP*.

5.2 RQ2. Which Factors Have an Effect on Participant Performance?

SV in Equation (4) features the instability of metrics like *MT*, *ER*, and *TP*. As Table 2 shows, *SV* is correlated with whether or not participants played games using a mouse and how long they used the mouse per day. High *SV* means the instability of session-level metrics, in other words, the lack of test-retest reliability. Regardless of the number of sessions in a study, the test-retest reliability can be made higher by enrolling participants who play games using a mouse or who have a long mouse usage time.

The test-retest reliability can be made higher if the factors that correlate with the metrics are made the same between all participants. For example, Table 1 suggests that the current hunger level (lower is more hungry) negatively correlated to *MT* and positively correlated to *ER*, which indicated that, as the participants were less hungry, they tended to accelerate the mouse-movement speed and thus the operations were more error-prone. Therefore, if we control the hunger level in each session, it could make the test-retest reliability higher. In the same manner, controlling the current sleepiness and current fatigue of the hand to hold the mouse could allow us to obtain the more stable performance.

Table 2 suggests that the current hunger level, busyness after the session, and whether or not games were played using a mouse all affect the performance. These factors affected the performance in multiple sessions by a single participant, and thus it is desirable to control them if researchers would like to obtain more stable task outcomes.

5.3 Limitation and Future Work

5.3.1 Definition of a new metric to evaluate test-retest reliability. In this paper, we recommended conducting experiments consisting of three or more sessions in order to observe stable performance. At the same time, the greater the number

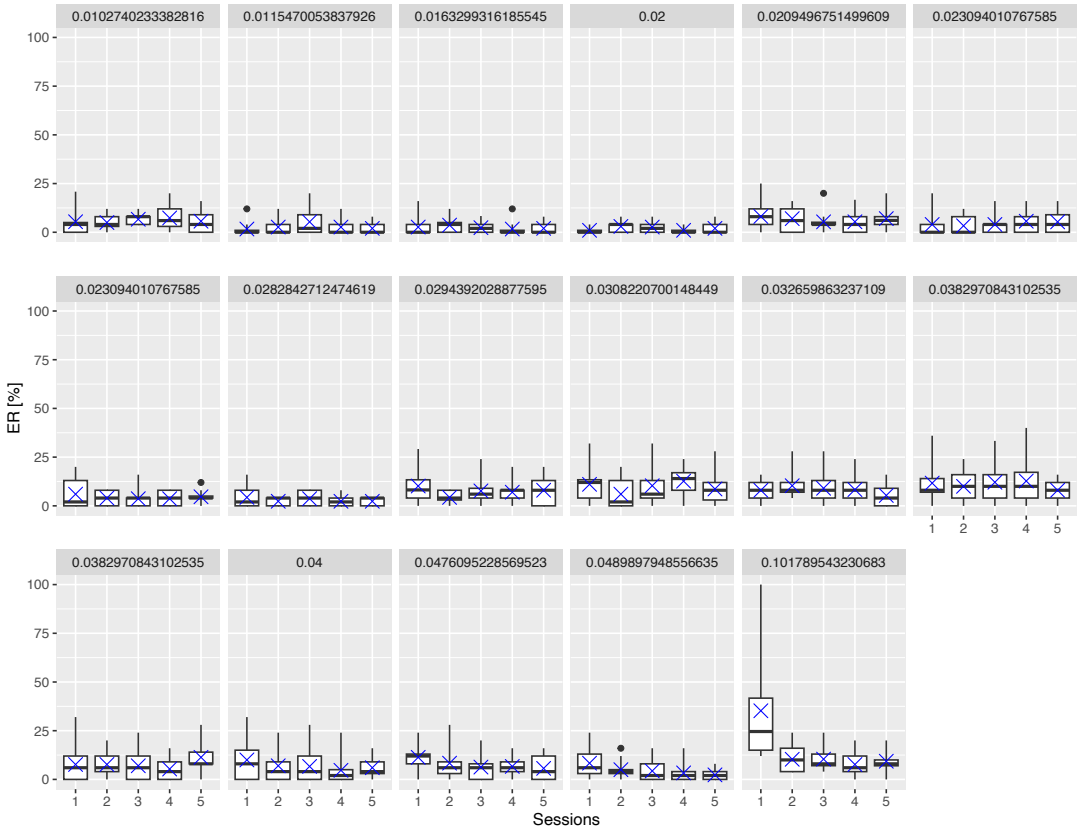


Fig. 6. *ER* of each session and participant. The title of each graph means *SV*.

of sessions, the fewer volunteer there are likely to be, and the more time, money, and effort are required on the part of researchers. Thus, we evaluated the results for the participant level to explore how to achieve lower *SV* in studies consisting of even a few sessions for higher test-retest reliability. However, we faced two issues. First, according to the current definition, a low *SV* (thus high test-retest reliability) does not necessarily mean that the participants do not improve their performance; see Fig. 1. Second, this experiment is constrained by a sample size of 17 participants, making its applicability to larger-scale experiments uncertain. Further experiments are necessary to verify whether three sessions would be sufficient in such larger-scale studies. If we derive a more appropriate metric and a prediction model for the number of required sessions, these will contribute significantly in the future.

5.3.2 Comprehensiveness of questionnaire items. The correlation between the performances and the questionnaire answers revealed which factors affect the performances and thus the test-retest reliability. We only had eight items in the session questionnaire and seven in the post-experiment one, but there are many other potentially significant factors such as mouse cursor speed, chair height, and exercise intensity, in addition to the resolution, size, and luminance of the display. It would be worthwhile to conduct an explorative study on the significant factors that participants should be asked about on questionnaires.

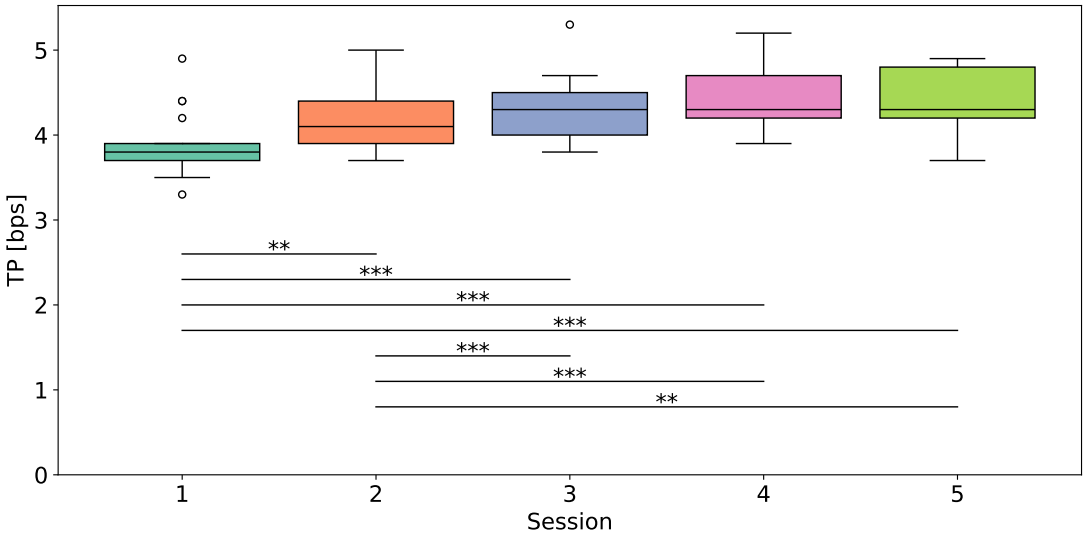


Fig. 7. *TP* of each session (**: $p < 0.01$, ***: $p < 0.001$).

Table 1. Correlation between participants' performance in each session and answers to the session questionnaire. We analyzed each participant's average performance in each session and the answers for all five days. Colored cells when correlation coefficients are -0.4 – -0.2 and 0.2 – 0.4.

Parameter	Sleeping hours last night	Quality of sleep last night (1: Could not sleep at all, 5: Slept well)	Current sleepiness (1: Very sleepy, 5: Not sleepy at all)	Current hunger level (1: Hungry, 5: Full)	Current eye fatigue (1: Very fatigued, 5: No fatigue)
MT	-0.0712	0.0188	-0.2053	-0.3775	-0.0251
ER	0.0215	0.1449	0.1818	0.2395	0.0826
TP	0.0413	-0.0202	0.0820	0.2463	-0.0662

Parameter	Current fatigue of the hand to hold the mouse (1: Very fatigued, 5: No fatigue)	Busyness of today before session (1: Very busy, 5: Not busy)	Busyness of today after session (1: Very busy, 5: Not busy)
MT	0.2681	-0.0910	-0.1223
ER	0.0500	0.0581	0.1482
TP	-0.2634	0.0814	0.0594

Table 2. Correlation between the average of participants' performance in each session and answers to the post-experiment questionnaire. We analyzed each participant's average performance in all sessions and the answers to the post-experiment questionnaire. Colored cells when correlation coefficients are -0.7 – -0.4, -0.4 – -0.2, 0.2 – 0.4, and 0.4 – 0.7.

Parameter	Age	Sex (Male: 1, Female: 0)	Dominant hand (Left: 1, Right: 0)	Dominant eye (Left: 1, Right: 0)	Whether or not playing games using a mouse (Yes: 1, No: 0)	Whether or not playing FPS (Yes: 1, No: 0)	How long using the mouse per day (hours)
MT	-0.0670	0.0341	-0.3985	0.1338	-0.1139	-0.1063	-0.2733
ER	-0.1889	0.1026	0.3481	-0.0595	0.0306	0.2066	0.1692
TP	0.1851	-0.1618	0.3281	-0.2466	0.1077	-0.0578	0.2122
SV (*MT*)	-0.1660	0.4332	-0.1722	0.0573	-0.5767	-0.2100	-0.4279
SV (*ER*)	0.0236	0.3380	-0.2114	0.0777	-0.3569	-0.1613	-0.2217
SV (*TP*)	0.1233	0.3403	-0.0342	0.2183	-0.5487	-0.0076	-0.2413

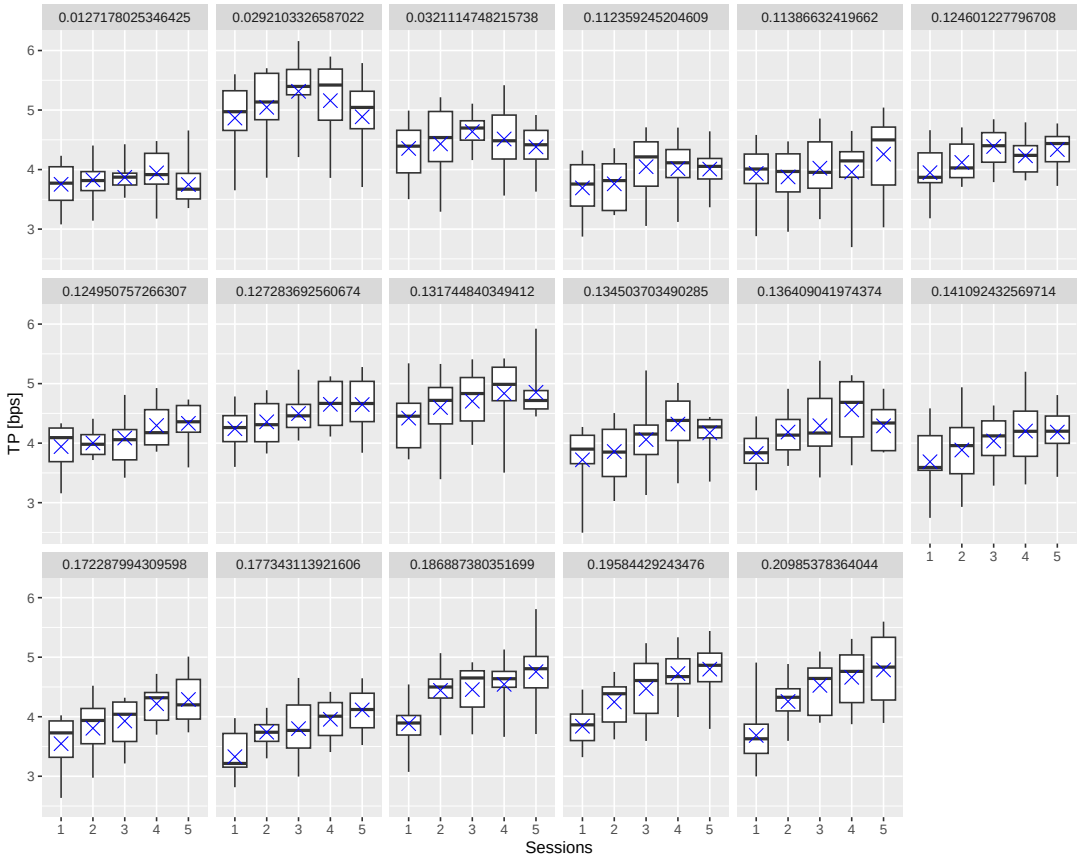


Fig. 8. *TP* of each session and participant. The title of each graph means *SV*.

6 CONCLUSION

We investigated test-retest reliability in a 2D-target pointing task by conducting an experiment consisting of five sessions with an interval of approximately one day. Our findings showed, that more than three sessions of experiments should be conducted in order to stabilize their performance on the 2D-target pointing task. We also found that the participants' current sleepiness, current hunger, and current fatigue of the hand to hold the mouse all affect the performance. In future work, we plan to define a model that can evaluate the improvement of participant' performance in terms of test-retest reliability. In addition, we will design a more detailed and itemized questionnaire to include additional items that may affect the performance of participants.

REFERENCES

[1] Nikola Banovic, Tovi Grossman, and George Fitzmaurice. 2013. 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* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1373–1382. https://doi.org/10.1145/2470654.2466181

[2] Israel Lucas Barros De Amorim, Suzane Santos Dos Santos, Ingrid Moreira Miranda Da Silva, Kamila Rios Da Hora Rodrigues, and Marcelle Pereira Mota. 2024. Gender Nuances in Human-Computer Interaction Research. In *Proceedings of the XXII Brazilian Symposium on Human*

Factors in Computing Systems (Maceió, Brazil) *(IHC '23)*. Association for Computing Machinery, New York, NY, USA, Article 54, 12 pages. https://doi.org/10.1145/3638067.3638077

[3] Xiaojun Bi, Yang Li, and Shumin Zhai. 2013. FFitts Law: Modeling Finger Touch with Fitts' Law. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. ACM, New York, NY, USA, 1363–1372. https://doi.org/10.1145/2470654.2466180

[4] Stuart K. Card, William K. English, and Betty J. Burr. 1978. Evaluation of Mouse, Rate-Controlled Isometric Joystick, Step Keys, and Text Keys for Text Selection on a CRT. *Ergonomics* 21, 8 (1978), 601–613. https://doi.org/10.1080/00140137808931762

[5] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a Replication Crisis in Empirical Computer Science. *Commun. ACM* 63, 8 (2020), 70–79. https://doi.org/10.1145/3360311

[6] Edward R.F.W. Crossman. 1956. The Speed and Accuracy of Simple Hand Movements. PH.d. Dissertation. University of Birmingham.

[7] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 754–768. https://doi.org/10.1145/3472749.3474784

[8] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39 (2007), 175–191.

[9] Leah Findlater, Joan Zhang, Jon E. Froehlich, and Karyn Moffatt. 2017. Differences in Crowdsourced vs. Lab-Based Mobile and Desktop Input Performance Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6813–6824. https://doi.org/10.1145/3025453.3025820

[10] Paul M. Fitts. 1954. The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement. *Journal of Experimental Psychology* 47, 6 (1954), 381–391. https://doi.org/10.1037/h0055392

[11] International Organization for Standardization. 2012. ISO/TS 9241-411:2012. https://www.iso.org/standard/54106.html. (accessed 2023-02-01).

[12] R. Chris Fraley and Brent W. Roberts. 2005. Patterns of Continuity: A Dynamic Model for Conceptualizing the Stability of Individual Differences in Psychological Constructs Across the Life Course. *Psychological Review* 112, 1 (2005), 60–74. https://doi.org/10.1037/0033-295X.112.1.60

[13] Timo Gnambs. 2014. A Meta-Analysis of Dependability Coefficients (Test–Retest Reliabilities) For Measures of the Big Five. *Journal of Research in Personality* 52 (2014), 20–28. https://doi.org/10.1016/j.jrp.2014.06.003

[14] Julien Gori, Olivier Rioul, and Yves Guiard. 2018. Speed-Accuracy Tradeoff: A Formal Information-Theoretic Transmission Scheme (FITTS). *ACM Trans. Comput.-Hum. Interact.* 25, 5, Article 27 (Sept. 2018), 33 pages. https://doi.org/10.1145/3231595

[15] Joy Paul Guilford. 1957. *Fundamental Statistics in Psychology and Education*. Vol. 41. McGraw-Hill Book Company, New York (330 West 42nd Street). 565 pages. https://doi.org/10.1002/sce.3730410357

[16] James J. Higgins and Suleiman Tashtoush. 1994. An aligned rank transform test for interaction. *Nonlinear World* 1, 2 (1994), 201 – 211.

[17] Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. 2014. Is Once Enough? On the Extent and Content of Replications in Human-Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3523–3532. https://doi.org/10.1145/2556288.2557004

[18] Jacob O. Wobbrock, Susumu Harada, Edward Cutrell, I. Scott MacKenzie. 2008. FittsStudy. https://depts.washington.edu/acelab/proj/fittsstudy/index.html. (accessed 2023-02-01).

[19] Alvin Jude, G. Michael Poor, and Darren Guinness. 2014. An Evaluation of Touchless Hand Gestural Interaction for Pointing Tasks with Preferred and Non-Preferred Hands. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (Helsinki, Finland) *(NordiCHI '14)*. Association for Computing Machinery, New York, NY, USA, 668–676. https://doi.org/10.1145/2639189.2641207

[20] Paul Kabbash, I. Scott MacKenzie, and William Buxton. 1993. Human Performance Using Computer Input Devices in the Preferred and Non-Preferred Hands. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 474–481. https://doi.org/10.1145/169059.169414

[21] I Scott MacKenzie. 1989. A Note on the Information-Theoretic Basis for Fitts' Law. *Journal of motor behavior* 21, 3 (1989), 323–330. https://doi.org/10.1080/00222895.1989.10735486

[22] I. Scott MacKenzie. 1992. Fitts' Law as a Research and Design Tool in Human-Computer Interaction. *Human-Computer Interaction* 7, 1 (march 1992), 91–139. https://doi.org/10.1207/s15327051hci0701_3

[23] I. Scott MacKenzie and Poika Isokoski. 2008. Fitts' Throughput and the Speed-Accuracy Tradeoff. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1633–1636. https://doi.org/10.1145/1357054.1357308

[24] Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. 2016. The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5421–5432. https://doi.org/10.1145/2858036.2858063

[25] Saaz Kaur Sahdra, Grant Duthie, Judy Kay, and Kalina Yacef. 2024. Monitoring Physical Health, Mental Health, Nutrition, and Sleep in Athletes to Improve Performance: Workshop Position Paper: Multimodal Sports Interaction: Wearables and HCI in Motion. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Melbourne VIC, Australia) *(UbiComp '24)*. Association for Computing Machinery, New York, NY, USA, 444–449. https://doi.org/10.1145/3675094.3678499

[26] K. C. Salter and R. F Fawcett. 1993. The art test of interaction: a robust and powerful rank test of interaction in factorial models. *Communications in Statistics - Simulation and Computation* 22, 1 (1993), 137–153. https://doi.org/10.1080/03610919308813085

[27] Shyam Prathish Sargunam, Kasra Rahimi Moghadam, Mohamed Suhail, and Eric D. Ragan. 2017. Guided Head Rotation and Amplified Head Rotation: Evaluating Semi-natural Travel and Viewing Techniques in Virtual Reality. In *2017 IEEE Virtual Reality (VR)*. 19–28. https://doi.org/10.1109/VR.2017.7892227

[28] James M. Schuerger, Karen L. Zarrella, and Annette S. Hotz. 1989. Factors That Influence the Temporal Stability of Personality by Questionnaire. *Journal of Personality and Social Psychology* 56, 5 (1989), 777–783. https://doi.org/10.1037/0022-3514.56.5.777

[29] Ather Sharif, Victoria Pao, Katharina Reinecke, and Jacob O. Wobbrock. 2020. The Reliability of Fitts's Law as a Movement Model for People with and without Limited Fine Motor Function. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 16, 15 pages. https://doi.org/10.1145/3373625.3416999

[30] R. William Soukoreff and I. Scott MacKenzie. 2004. Towards a Standard for Pointing Device Evaluation, Perspectives on 27 Years of Fitts' Law Research in HCI. *Int. J. Hum.-Comput. Stud.* 61, 6 (2004), 751–789. https://doi.org/10.1016/j.ijhcs.2004.09.001

[31] R. William Soukoreff and I. Scott MacKenzie. 2004. Towards a Standard for Pointing Device Evaluation, Perspectives on 27 Years of Fitts' Law Research in HCI. *International Journal of Human-Computer Studies* 61, 6 (2004), 751–789. https://doi.org/10.1016/j.ijhcs.2004.09.001 Fitts' law 50 years later: applications and contributions from human-computer interaction.

[32] John Vines, Gary Pritchard, Peter Wright, Patrick Olivier, and Katie Brittain. 2015. An Age-Old Problem: Examining the Discourses of Ageing in HCI and Strategies for Future Research. *ACM Transactions on Computer-Human Interaction* 22, 1, Article 2 (Feb. 2015), 27 pages. https://doi.org/10.1145/2696867

[33] Chockalingam Viswesvaran and Deniz S. Ones. 2000. Measurement Error in "Big Five Factors" Personality Assessment: Reliability Generalization across Studies and Measures. *Educational and Psychological Measurement* 60, 2 (2000), 224–235. https://doi.org/10.1177/00131640021970475

[34] Max L. Wilson, Ed H. Chi, Stuart Reeves, and David Coyle. 2014. RepliCHI: The Workshop II. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI EA '14)*. Association for Computing Machinery, New York, NY, USA, 33–36. https://doi.org/10.1145/2559206.2559233

[35] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[36] Jacob O. Wobbrock, Kristen Shinohara, and Alex Jansen. 2011. The Effects of Task Dimensionality, Endpoint Deviation, Throughput Calculation, and Experiment Design on Pointing Measures and Models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1639–1648. https://doi.org/10.1145/1978942.1979181

[37] Shota Yamanaka. 2022. Test-Retest Reliability on Movement Times and Error Rates in Target Pointing. In *Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 178–188. https://doi.org/10.1145/3532106.3533450

[38] Shota Yamanaka and Hiroki Usuba. 2020. Rethinking the Dual Gaussian Distribution Model for Predicting Touch Accuracy in On-Screen-Start Pointing Tasks. *Proc. ACM Hum.-Comput. Interact.* 4, ISS, Article 205 (Nov. 2020), 20 pages. https://doi.org/10.1145/3427333