Vol.2022-HCI-199 No.6
2022/8/22

2次元ポインティングタスクにおける再テスト信頼性の予備調査

加藤 進吾[1,a)]　横山 海青[1]　日高 拓真[1]　佐藤 大樹[2]　山中 祥太[3,b)]　志築 文太郎[4,c)]

概要：再テスト信頼性とは，同一参加者が 2 つのセッションにおいて同一タスクを実行した際，両セッションの結果がどの程度同じになるかを問う度合いである．再テスト信頼性を調査することにより，実験条件外の要因による影響を知ることができ，これに伴って研究者は実験から得られた結論の信頼性を検討できる．先行研究において，1 次元ポインティングタスクの再テスト信頼性が低いことが示されているが，2 次元ポインティングタスクについては調査されていない．そこで，本研究では 2 次元ポインティングタスクにおけるマウスの移動時間の観点から，参加者がパフォーマンスをどの程度維持できるかを調査した．この結果，セッション間でマウスの移動時間にばらつきがある可能性を確認した．このことから，2 次元ポインティングタスクにおける再テスト信頼性が低い可能性が示唆された．

1. はじめに

再テスト信頼性とは，同一参加者が 2 つのセッションにおいて同一のタスクを行った際，両セッションの結果がどの程度同じになるかを問う度合い [1] である．HCI 分野においては，1 回限りのユーザ実験の結果を統計解析して結論を導くことも多いが，偶然に高い（あるいは低い）測定結果が得られることもありうる．再テスト信頼性を調査することにより，実験条件外の要因による影響を知ることができ，これに伴って研究者は実験から得られた結論の信頼性を検討できる．1937 年ごろから注目された再テスト信頼性 [2] は，特に心理学分野において調査されている [3–6]．例えば，同一参加者群が数週間後に同一のアンケートを与えられた際，同一の回答が得られるかを調査した研究 [5] がなされている．

HCI 分野において，実験の再現性を重要視する研究は多く存在する（例：[7–9]）．一方，再テスト信頼性を調査した研究は少ない．再現性においては，研究者は，実験条件（装置，参加者の属性，および課題など）を同一にすることに着目する．一方，再テスト信頼性においては，同一の実験条件に加えて，同一の参加者であることおよび同一の実験結果が得られることにも着目する必要がある．なお，再テスト信頼性を調査する際，同一参加者が 2 つのセッション間において同一の結果を出すことが理想的である．一方，心理学分野において，同一の結果が出ない例が複数報告されている [3–6]．

HCI 分野において再テスト信頼性を調査した研究として，Sharif ら [10] および山中 [1] の研究が挙げられる．両研究とも，1 次元ポインティングタスクの再テスト信頼性の調査を行い，結果として再テスト信頼性が低いことを示した．2 次元ポインティングタスクについても，複数セッションによる実験は行われている [11] が，再テスト信頼性の調査までは行われていない．

本研究において，我々は，2 次元ポインティングタスク，具体的には ISO9241-411 [12] に示される multi-directional pointing task におけるマウスの移動時間（MT）の観点から再テスト信頼性の予備的な調査を行った．具体的には，著者 4 人を参加者として，同一条件における実験を 5 日間行った．この結果，1 日目，2–4 日目，および 5 日目の間において，MT にばらつきがある可能性が示された．このことから，2 次元ポインティングタスクにおいても再テスト信頼性が低い可能性が示唆された．本研究の貢献は，2 次元ポインティングタスクにおける再テスト信頼性を初めて調査し，結果として Sharif らおよび山中の結論を 2 次元ポインティングタスクに拡張できる可能性を示したことである．

2. Fitts の法則および ID の定義

本研究では，実験結果の Fitts の法則に対するモデル適

[1] 筑波大学 情報理工学位プログラム
[2] 筑波大学 情報メディア創成学類
[3] ヤフー株式会社
[4] 筑波大学 システム情報系
[a)] skato@iplab.cs.tsukuba.ac.jp
[b)] syamanak@yahoo-corp.jp
[c)] shizuki@cs.tsukuba.ac.jp

合性に基づいて再テスト信頼性を調査する．Fitts の法則とは，マウスなどの入力装置を用いてターゲットを選択する所要時間のモデルである [13]．今回，我々が用いる Fitts の法則は，次式で定義されるものである [14]．

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

なお，MT はマウスの移動時間を表す．また，a および b は環境に依存する定数である．D はターゲットまでの距離，W はターゲットの幅を表す．なお，ID（*Index of Difficulty*）はターゲット選択の難易度を表す．

3. 関連研究

本節では，Fitts の法則の再テスト信頼性に関連する研究として，1 次元ポインティングタスクの再テスト信頼性を調査した研究，および操作手法の長期実験を行った研究を示す．

3.1 1 次元ポインティングタスクの再テスト信頼性を調査した研究

1 次元ポインティングタスクの再テスト信頼性を調査した研究として，Sharif らの研究 [10] および山中の研究 [1] が挙げられる．Sharif ら [10] は，Fitts の法則に対するモデル適合性の再テスト信頼性を調査するために，異なるエラー率（ER）において 1 次元ポインティングタスクのスループット（TP）を計測した．この計測の結果，各参加者について，セッション間において TP が異なる参加者および安定した参加者がいたこと，参加者全体の TP の平均がセッション間において大幅に変化したことから，再テスト信頼性が低いとした．また，山中 [1] は，再テスト信頼性の調査のために，1 次元ポインティングタスクにおける MT，ER，および TP を計測した．この計測の結果，一部の参加者の MT，ER，および TP がセッション間において大幅に変化したことから，再テスト信頼性が低いとした．

本研究においては，2 次元ポインティングタスクにおける再テスト信頼性を調査する．また，山中の研究 [1] ではクラウドソーシングを用いて実験を行っている．実験を現地で行う場合と遠隔で行う場合とでは，後者の方がタスクの精度が低下するという結果が報告されている [15]．そのため，我々は現地での実験を行う．

3.2 操作手法の長期実験を行った研究

操作手法の長期実験を行った研究として，提案した新手法の性能を既存手法と比較したものがある [16, 17]．Harada ら [16] は，Vocal Joystick と呼ばれる，音声をもとにマウスの移動方向を決定する操作手法を用いた長期実験を行った．なお，実験は 2.5 週間に渡って行われた．この実験の結果，セッションを重ねるごとにカーソルの移動時間が少なくとも 20%以上小さくなることが確認された．また，Mahmud ら [17] は，2 種類の音声カーソル操作システムを提案し，参加者に対して 5 日間の実験を行った．結果より，実験セッション数が進むにつれて操作が上達していることが分かった．一方，この結果は，音声カーソル操作システムにおける再テスト信頼性が低いことを示唆する．

これらの研究は，長期実験を行っている一方で，同条件および同参加者の実験，つまり再テスト信頼性を調査するための実験を行っていない．このため，実験の結果が新手法および既存手法の差であるのか，参加者の慣れなどの実験条件外の影響であるかが確かめられていない．我々の研究では，2 次元ポインティングにおける実験結果が実験条件による影響であるかを検討できるようにするため，調査を行う．

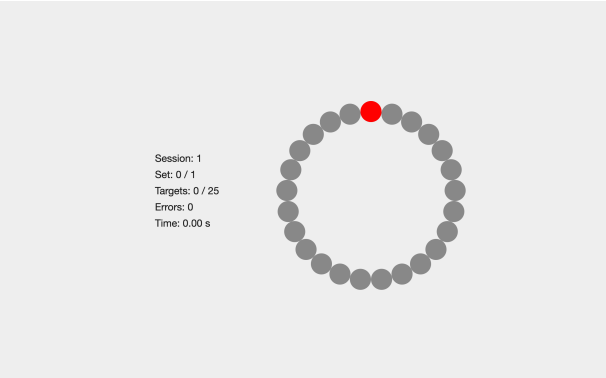

図 1 実験アプリケーションの表示例．

4. 著者実験

著者を参加者とした実験（以降，著者実験）を行った．参加者は 5 日間連続して，毎日決まった時間帯に 2 次元ポインティングタスクを行った．MT が測定の対象となった．

4.1 参加者

22–23（M=22.8, SD=0.50）歳の著者 4 名が実験に参加した．全員が右利きであった．また，全員が男性であった．

4.2 実験装置

実験アプリケーション（図 1）はラップトップ PC（iiyama, STYLE-15FX160-i7-RASX）上にて実行された．なお，ラップトップ PC のディスプレイの解像度は 1920 × 1080 であり，大きさは 15.6 型である．また，入力装置としてマウス（BUFFALO, BSMBU05）をラップトップ PC に接続した．机の表面の凹凸によるマウス操作への影響を軽減するため，参加者は 430 mm × 290 mm のマウスパッド（Logicool G, G640r）上にてマウスを使用した．なお，このサイズは参加者がタスクを行う際にマウスのクラッチを必要としないサイズである．

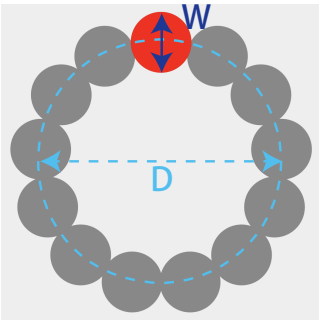

図 2 ターゲットの幅 W およびターゲット間の距離 D の定義．ターゲットの幅 W は，ターゲットとなる円の直径である．すべてのターゲットは，その中心を直径 D の円に重ねつつ，等間隔に並ぶ．

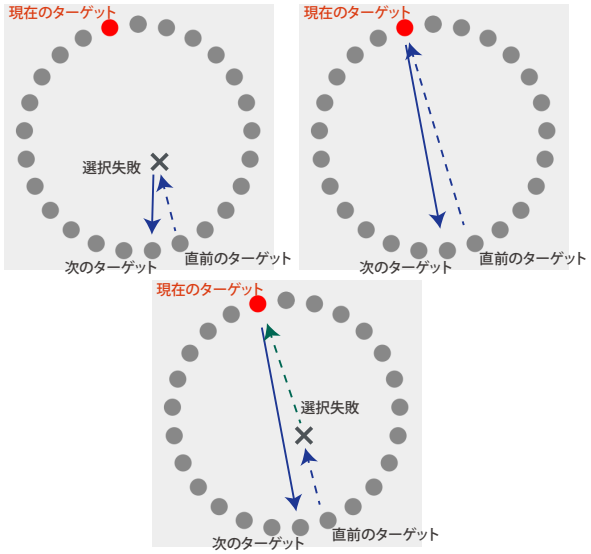

図 3 現在のターゲットの選択に失敗した際，現在のターゲットの選択をやり直さずに次のターゲットの選択に移る（左上）と，次のターゲットを選択する際のカーソルの移動距離および開始地点（青実線矢印）が，現在のターゲットの選択に成功した際（右上）とは異なることになる．一方，現在のターゲットの選択をやり直せば（下），同じ移動距離および開始地点から次のターゲットの選択を行うことができる．

4.3 タスク

参加者は実験用アプリケーションを用いて，multi-directional pointing task を行った．参加者は実験用アプリケーション上の赤色のターゲットをクリックにより選択する．ターゲットの選択に成功すると，次のターゲットが赤色にて示される．なお，ターゲットの選択に失敗した場合はクリックすべきターゲットの色が 2 回オレンジ色（$R = 255, G = 153, B = 102$）に点滅することにより，参加者に選択の失敗を知らせる．また，参加者はターゲットの選択に成功するまで，同じターゲットの選択を続ける必要がある．

4.4 実験用アプリケーション

我々の実験用アプリケーションは FittsStudy [18] に基づく．FittsStudy は，ISO9241-411 [12] に従って 1 次元および 2 次元ポインティングスタディを実施，および分析するためのツールである．本ツールではターゲット間距離 D およびターゲット幅 W を設定できる．2 次元ポインティングタスクにおけるターゲット間距離 D およびターゲット幅 W の定義を図 2に示す．

しかし，図 3に示す通り，FittsStudy を用いた 2 次元ポインティングタスクにおいては，ターゲットの選択の成否に関わらず，クリック操作によって次のターゲットの選択へと移る．すなわち，現在のターゲットの選択に失敗したとき，次のターゲットへの距離が D とは異なる可能性がある．そこで，我々は独自の実験用アプリケーションを作成した．本実験用アプリケーションでは，先行研究 [19,20] と同様に，ターゲットの選択に失敗した場合に参加者は次のターゲットの選択に移るのではなく，ターゲットの選択をやり直すよう実装されている．これにより，FittsStudy に比べて次のターゲットへの距離がより常に D に近くなる．

4.5 実験手順

参加者は 1 日に 1 セッション実験を行い，それを 5 日間繰り返した．また，2 セッション目以降の開始時刻を，1 セッション目の開始時刻から前後 4 時間の範囲とすることによって．決まった時間帯に実験を行うようにした．

全てのセッションにおいて，参加者はまずポインティングタスクの練習を行い，続いて計測を行った．

計測に用いた W は 4 通り（8，20，38，78 pixel）であり，また D は 3 通り（300，440，630 pixel）であった．つまり，タスクは計 12 種類となる．なお，W および D は，広い範囲（2.28–6.32 bits）の ID におけるポインティングが出題されるよう定められた．

ポインティングパフォーマンスの傾向を高精度に測定するためには，各タスクに対して 15–25 回のクリックを行わせることが推奨されている [19]．そのため，参加者は 1 セッションあたり 12 種類のタスクそれぞれに対して 25 回のクリックを行った．

なお，出題順が実験結果へ与える影響を低減するため，参加者内および参加者間において出題順を同一にした．また，参加者が実験において使用するマウスに慣れている可能性は低い．このため， ID を次第に大きくすることによって，参加者がマウスに慣れつつポインティングの難易度を上げられると考え，出題順を ID の昇順，すなわちポインティングが容易な順とした．

各セッション開始前には練習タスクを設けた．なお，計測の最初に出題される W と D の組み合わせ（ID= 2.28 bits）よりも簡単な組み合わせを練習タスクに用いる場合，ターゲットが過度にクリックしやすくなり，参加者が視覚フィードバックに頼ることなくタスクを行える可能性があるため，練習の効果が小さくなる可能性がある．したがって，

先行研究 [1] に倣って，練習タスクを $W = 30\,\text{pixel}$ かつ $D = 400\,\text{pixel}$（ID= 3.84 bits）という 2.28–6.32 bits の中程度の ID における 25 回のターゲット選択とした．

ID は 12 種類，クリック数は最小で 25，参加者は 4 名であったため，収集したデータ数は各セッションにつき $12 \times 25 \times 4 = 1200$ 個となった．参加者は 1 セッションを 15 分以内に完了することができた．

5. 実験結果および考察

著者実験の結果，および結果の考察を行う．

5.1 Fitts の法則のモデルとの適合度

本節において，著者実験の結果に対する，Fitts の法則のモデルの適合度を示す．本実験における参加者は，いずれもマウスの操作経験が十分であるため，新手法の調査においてよくみられる，新手法への慣れによる影響はないものと考えられる．なお，文章中に出てくる補正 R^2， AIC，および BIC とはモデルへの適合度を示す値である．補正 R^2 は値が高いほど適合度が高く， AIC および BIC は値が低いほど適合度が高いことを示す．

図 4は，各セッションの結果を $MT = a + b \times ID$ に線形回帰した結果である．いずれのセッションにおいても，線形回帰の補正 R^2 が 0.9 以上あることから，実験の結果は Fitts の法則に適合しているといえる．

5.2 MT

各セッションにおけるパフォーマンスとして MT を比較した．表 1に，各セッションにおける MT を示す．また，セッションごとの MT の変化を可視化するため，図 5に各セッションにおける MT の平均を示す．

AIC は，1 セッション目，2–4 セッション目，および 5 セッション目において，それぞれ 2 以上の差があることから，これらのセッションの間において，有意差があることが分かった．また，各セッションの BIC は，1 セッション目，2–4 セッション目，および 5 セッション目において，それぞれ 2 以上の差があることから，これらのセッションの間においても，有意差があることが分かった．したがって，1 セッション目，2–4 セッション目，5 セッション目は互いにそれぞれ Fitts の法則の適合度が異なり，より遅い実験日のデータほど MT の予測精度が高いといえる．また，これらから，2 次元ポインティングタスクの再テスト信頼性が低い可能性が示唆される．

6. 本研究の制約および今後の計画

本節においては，本研究の制約および今後の計画を述べる．

6.1 TP の算出

本実験において，実験アプリケーションの不具合により，参加者がミスクリックをした際のクリック座標が記録できなかったため，TP の算出ができなかった．今後の実験においては，不具合を修正した実験アプリケーションを用いて， TP を算出する．また，算出した TP を用いて，新たな考察および Sharif らおよび山中の結果との比較を行う予定である．

6.2 参加者間におけるタスクを実行する時間帯の統一

参加者が実験を行う時間帯によって，結果が変わる可能性が考えられる．例えば，参加者によっては，午前中に実験を行った方が集中力が高くなり，夜中に行うよりも結果が向上することが考えられる．今回の実験においては，参加者間で実験を行う時間帯を統一していない．そのため，今後の実験では，参加者間における実験の時間帯を統一した場合に結果がどのように変化するのか調査する予定である．

6.3 クリック判定のタイミングの変更

本実験に用いられた実験用アプリケーションは，マウスのボタンが押された状態から離された状態になったとき，クリックが行われたと判定している．しかし，FittsStudy [18] は，マウスのボタンが離された状態から押された状態になったとき，クリックが行われたと判定している．したがって，クリック判定のタイミングを FittsStudy [18] と同じタイミングに変更する必要がある．

6.4 アンケート調査と実験結果の相関分析

今回は，参加者の性別，利き手，およびマウスの操作経験についての情報を，アンケート調査により収集した．本研究の目的は，実験条件外の要因が実験に及ぼす影響の程度を調査することである．そのため，「実験前の予定」，「実験後の予定」，および「その日の体調」など，アンケート調査の結果と実験結果の相関を分析することも必要であると考える．

6.5 実験タスクにおける順序のランダマイズ

本実験では，実験タスクの出題順を ID の昇順，すなわちポインティングが容易な順として統一した．これは出題順が実験結果へ与える影響を低減するためである．一方で，出題順を固定することにより，実験終盤になるにつれて疲労度や慣れの影響が生じている可能性も考えられる．そのため，今後は実験タスクの出題順をランダマイズすることでカウンターバランスをとり，そのうえで実験結果がどのように変化するのかを調査することも重要であると考えられる．

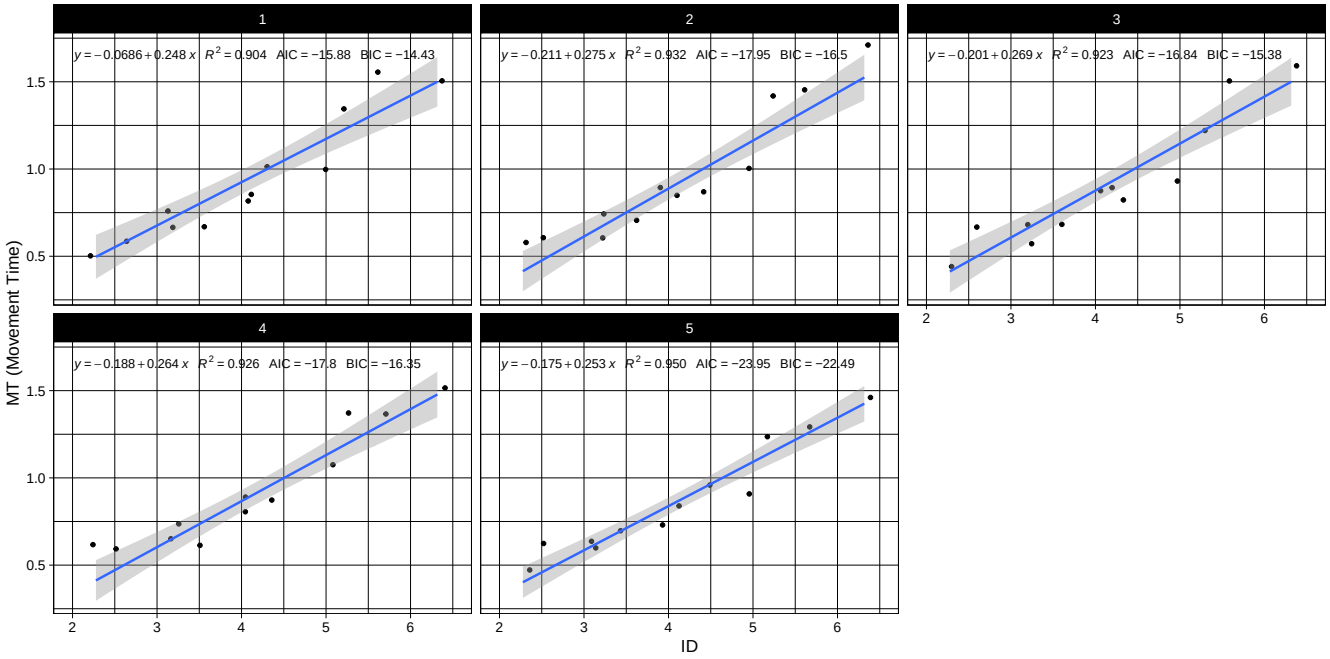

図 4 各セッションにおける *ID* ごとの *MT* の平均．回帰分析には全員分の結果の平均 *MT* を用いた．適合度を示す補正 R^2 は値が高いほど適合度が高く，*AIC* および *BIC* は値が低いほど適合度が高いことを示す．また，95%信頼区間は灰色の領域である．

表 1 各セッションにおける各参加者の *MT* の平均．カラム名に含まれる数字は分析対象のセッションを表す．つまり，1 セッション目の *MT* の平均は「平均_1」となる．

参加者	平均_1	標準偏差_1	平均_2	標準偏差_2	平均_3	標準偏差_3	平均_4	標準偏差_4	平均_5	標準偏差_5
1	0.94	0.34	0.90	0.36	0.92	0.36	0.90	0.41	0.87	0.36
2	0.88	0.31	0.86	0.33	0.83	0.34	0.84	0.34	0.81	0.30
3	0.99	0.39	0.94	0.43	0.96	0.43	0.91	0.36	0.89	0.35
4	1.02	0.40	0.98	0.43	0.94	0.41	0.94	0.38	0.92	0.41

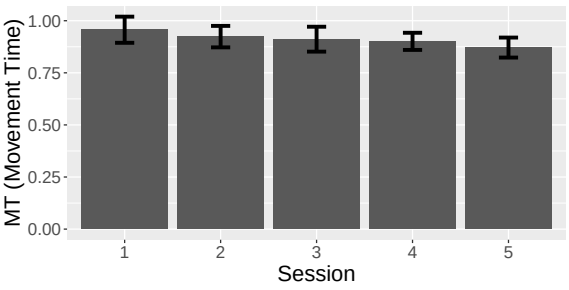

図 5 各セッションにおける *MT* の平均．

7. おわりに

本稿では，*MT* の観点から 2 次元ポインティングタスクにおける再テスト信頼性の予備調査を 5 日間にわたって行った．結果より，1 日目，2–4 日目，および 5 日目の間においてパフォーマンスに有意差がみられたことから，2D ポインティングタスクにおいても再テスト信頼性が低いことが示唆された．これにより，Sharif らおよび山中の結論を 2 次元ポインティングタスクに拡張できる可能性を示した．今後は，6 節で述べた内容を実装したうえで，著者以外を参加者とした実験を行う．

参考文献

[1] Yamanaka, S.: Test-Retest Reliability on Movement Times and Error Rates in Target Pointing, *Designing Interactive Systems Conference*, DIS '22, New York, NY, USA, Association for Computing Machinery, pp. 178–188 (online), DOI: 10.1145/3532106.3533450 (2022).

[2] Kuder, G. and Richardson, M.: The theory of the estimation of test reliability, *Psychometrika*, Vol. 2, No. 3, pp. 151–160 (online), DOI: 10.1007/BF02288391 (1937).

[3] Fraley, R. C. and Roberts, B. W.: Patterns of Continuity: A Dynamic Model for Conceptualizing the Stability of Individual Differences in Psychological Constructs Across the Life Course., *Psychological Review*, Vol. 112, No. 1, pp. 60–74 (online), DOI: 10.1037/0033-295X.112.1.60 (2005).

[4] Gnambs, T.: A Meta-Analysis of Dependability Coefficients (Test–Retest Reliabilities) For Measures of the Big Five, *Journal of Research in Personality*, Vol. 52, pp. 20–28 (online), DOI: 10.1016/j.jrp.2014.06.003 (2014).

[5] Schuerger, J. M., Zarrella, K. L. and Hotz, A. S.: Factors That Influence the Temporal Stability of Personality by Questionnaire, *Journal of Personality and Social*

Psychology, Vol. 56, No. 5, pp. 777–783 (online), DOI: 10.1037/0022-3514.56.5.777 (1989).

[6] Viswesvaran, C. and Ones, D. S.: Measurement Error in "Big Five Factors" Personality Assessment: Reliability Generalization across Studies and Measures, *Educational and Psychological Measurement*, Vol. 60, No. 2, pp. 224–235 (online), DOI: 10.1177/00131640021970475 (2000).

[7] Cockburn, A., Dragicevic, P., Besançon, L. and Gutwin, C.: Threats of a Replication Crisis in Empirical Computer Science, *Commun. ACM*, Vol. 63, No. 8, pp. 70–79 (online), DOI: 10.1145/3360311 (2020).

[8] Hornbæk, K., Sander, S. S., Bargas-Avila, J. A. and Grue Simonsen, J.: Is Once Enough? On the Extent and Content of Replications in Human-Computer Interaction, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, New York, NY, USA, Association for Computing Machinery, pp. 3523–3532 (online), DOI: 10.1145/2556288.2557004 (2014).

[9] Wilson, M. L., Chi, E. H., Reeves, S. and Coyle, D.: RepliCHI: The Workshop II, *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, New York, NY, USA, Association for Computing Machinery, pp. 33–36 (online), DOI: 10.1145/2559206.2559233 (2014).

[10] Sharif, A., Pao, V., Reinecke, K. and Wobbrock, J. O.: The Reliability of Fitts's Law as a Movement Model for People with and without Limited Fine Motor Function, *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3373625.3416999 (2020).

[11] Jude, A., Poor, G. M. and Guinness, D.: An Evaluation of Touchless Hand Gestural Interaction for Pointing Tasks with Preferred and Non-Preferred Hands, *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, NordiCHI '14, New York, NY, USA, Association for Computing Machinery, pp. 668–676 (online), DOI: 10.1145/2639189.2641207 (2014).

[12] International Organization for Standardization: ISO/TS 9241-411:2012, https://www.iso.org/standard/54106.html. (accessed 2022-07-19).

[13] Fitts, P. M.: The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement., *Journal of Experimental Psychology*, Vol. 47, No. 6, pp. 381–391 (online), DOI: 10.1037/h0055392 (1954).

[14] MacKenzie, I. S.: A Note on the Information-Theoretic Basis for Fitts' Law, *Journal of motor behavior*, Vol. 21, No. 3, pp. 323–330 (online), DOI: 10.1080/00222895.1989.10735486 (1989).

[15] Findlater, L., Zhang, J., Froehlich, J. E. and Moffatt, K.: Differences in Crowdsourced vs. Lab-Based Mobile and Desktop Input Performance Data, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, New York, NY, USA, Association for Computing Machinery, pp. 6813–6824 (online), DOI: 10.1145/3025453.3025820 (2017).

[16] Harada, S., Wobbrock, J. O., Malkin, J., Bilmes, J. A. and Landay, J. A.: Longitudinal Study of People Learning to Use Continuous Voice-Based Cursor Control, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, New York, NY, USA, Association for Computing Machinery, pp. 347–356 (online), DOI: 10.1145/1518701.1518757 (2009).

[17] Mahmud, M., Sporka, A. J., Kurniawan, S. H. and Slavík, P.: A Comparative Longitudinal Study of Non-verbal Mouse Pointer, *Human-Computer Interaction – INTERACT 2007* (Baranauskas, C., Palanque, P., Abascal, J. and Barbosa, S. D. J., eds.), Springer Berlin Heidelberg (2007).

[18] Jacob O. Wobbrock, Susumu Harada, Edward Cutrell, I. Scott MacKenzie: FittsStudy, https://depts.washington.edu/acelab/proj/fittsstudy/index.html. (accessed 2022-07-01).

[19] Soukoreff, R. W. and MacKenzie, I. S.: Towards a Standard for Pointing Device Evaluation, Perspectives on 27 Years of Fitts' Law Research in HCI, *Int. J. Hum.-Comput. Stud.*, Vol. 61, No. 6, pp. 751–789 (online), DOI: 10.1016/j.ijhcs.2004.09.001 (2004).

[20] MacKenzie, I. S.: Fitts' Law as a Research and Design Tool in Human-Computer Interaction, *Hum.-Comput. Interact.*, Vol. 7, No. 1, pp. 91–139 (online), DOI: 10.1207/s15327051hci0701_3 (1992).