Vol.2025-HCI-215 No.23
Vol.2025-UBI-88 No.23
2025/11/27

クラウドソーシングを活用したGUI実験における参加者スクリーニング手法のスマートフォンでの検証

三山 貴也[1]　中村 聡史[1]　山中 祥太[2]

概要：クラウドソーシングを活用したGUI実験では，多くの参加者を素早く募集でき，UI操作時のデータを大量に収集できる．一方で，指示を守らない参加者や雑な操作を行う参加者も存在するため，実験データの品質が低下する懸念がある．我々はこれまで，実験前の事前タスクを利用した参加者のスクリーニング手法について検証を行い，不適切な参加者の割合を低くするほどGUI操作のパフォーマンスモデルの適合度が向上することを明らかにしてきた．しかしこれまでの研究はPC環境での検証にとどまり，画面表示の統制が不十分で実験の条件数も限られていたため，厳密な検証ができていなかった．そこで本稿では，スマートフォン環境で視覚刺激の表示をmm単位で統制し，条件数を増やした実験を通じてスクリーニング手法の有用性を厳密に検証した．その結果，不適切な参加者の割合を低くするほどモデル適合度が向上する傾向が再確認され，事前タスクを利用したスクリーニング手法の有用性がより明確に示唆された．

1. はじめに

グラフィカルユーザインタフェース（GUI）に関するユーザ実験において，クラウドソーシングの利用が一般的になっている．利点の一つとして，GUI操作時のデータを大量に収集できることから，ポインティングタスクにおけるエラー率のような発生確率の低い事象を適切に扱えることが報告されている [1]．一方，クラウドソーシング実験の参加者は実験室実験の参加者に比べてポインティングタスクの操作が不正確な傾向があり，2倍以上のエラー率が観察されたという報告も存在する [2]．このように，クラウドソーシングを活用したGUI実験において，募集された参加者の性質が実験結果に影響を及ぼし，実験データの品質を低下させる懸念がある．

我々はこれまでの研究 [3, 4] において，クラウドソーシングを活用したGUI実験におけるデータ品質向上を目的として，実験前の事前タスクによって参加者のスクリーニングを行い，適切なユーザ群のみに本来目的とする実験を依頼するアプローチを提案してきた．ここでは，事前タスクとして簡単な操作課題である画像のリサイズ，実験の主タスクとしてポインティングタスクを実施した結果，事前タスクの結果が不適切な参加者の割合を低くするほどGUI操作のパフォーマンスモデルの適合度が向上することを明らかにしている．しかしこれまでの研究はPC環境での検証にとどまり，画面表示の統制が不十分で実験の条件数も限られていたため，厳密な検証ができていなかった．

そこで本稿では，スマートフォン環境で視覚刺激の表示をmm単位で統制し，条件数を増やした実験を通じてスクリーニング手法の有用性を厳密に検証する．事前タスクと主タスクの両方について厳密な評価を行い，これまでの研究で得られた結果の傾向を再確認することで，スクリーニング手法の有用性をより明確に示すことを目的とする．

[1] 明治大学
Meiji University
[2] LINE ヤフー株式会社
LY Corporation

2. 関連研究

2.1 クラウドソーシング実験のデータ品質

クラウドソーシング実験のデータ品質に影響を与える要因として，不注意な参加者の存在が指摘されており，参加者のうち45.9%が何らかの不注意な行動をしたという報告 [5] がある．これに対し，不注意な回答の検出には反応時間や自由記述の分析が効果的だとされている [6]．またOppenheimerら [7] は参加者が指示を読んでいるかを測定する質問（Instructional Manipulation Check，IMC）を提案し，IMCを通過した参加者群と通過しなかった参加者群では異なる実験結果になる可能性があるとしている．

以上のように，クラウドソーシング実験のデータ品質に関する検証や対策が行われている．本研究は，実験前の事前タスクによって参加者のスクリーニングを行うアプローチにより，これまでの取り組みをさらに改善する可能性を探るものである．

2.2 クラウドソーシングを活用した GUI 実験

GUI 実験について，クラウドソーシング実験と実験室実験の結果の比較が行われている．Komarov ら [8] は Bubble Cursor [9] が従来のカーソルよりも操作時間を短縮させるという結果が，クラウドソーシング実験でも再現されたと報告している．また Findlater ら [2] は，マウスおよびタッチ操作のポインティングタスクにおいて，クラウドソーシング実験の参加者は実験室実験の参加者よりも操作時間が短くエラー率が高いため，正確さよりも速さを重視する傾向があるとしている．

以上のように，クラウドソーシングを活用した GUI 実験においてもデータ品質に関する検証が行われている．本研究では，GUI 実験特有の操作やインタラクションに適した事前タスクによって参加者をスクリーニングすることで，さらなるデータ品質の向上を目指す．

2.3 ポインティングタスク

ポインティングタスクの代表的なモデルであるフィッツの法則は，ターゲットまでの距離 A とターゲットの幅 W に基づいて，最初のクリックまでの時間 MT を予測できる [10, 11]．また，ターゲットの範囲外をクリックするエラー率 ER に関して，操作が速くなると ER が増加し，慎重になると減少することが知られている [12, 13]．さらに，ER を予測するモデルも提案されており，長方形や円形のターゲットについてモデルが存在する [14, 15]．

フィッツの法則に関する実験では，できるだけ速く正確に操作するように指示することが一般的だが，参加者によって速さと正確さのバランスにバイアスがあることも知られている [16, 17]．また，バイアスがかかった状態ごとにタスクを行うことで，速さ重視の場合に MT が減少する一方で ER が増加するといったように，状況に応じた操作を評価できるという報告もある [18]．

3. 提案アプローチ

3.1 コンセプト

本研究では，クラウドソーシングを活用した GUI 実験におけるデータ品質向上を目的として，実験前の事前タスクによって参加者のスクリーニングを行い，本来目的とする実験（主タスク）を適切なユーザ群のみに依頼するアプローチを提案する．具体的には，主タスクの操作と関連のある操作を含む事前タスクを設け，その操作パフォーマンスに基づいて適切なユーザを抽出することで，事前タスクのみの実施によって適切なユーザ群に主タスクの実験を依頼可能とする．これにより，参加者の多くが適切な操作を行うユーザとなり，実験データの品質向上が期待できると考える．提案アプローチのイメージ図を図 1 に示す．

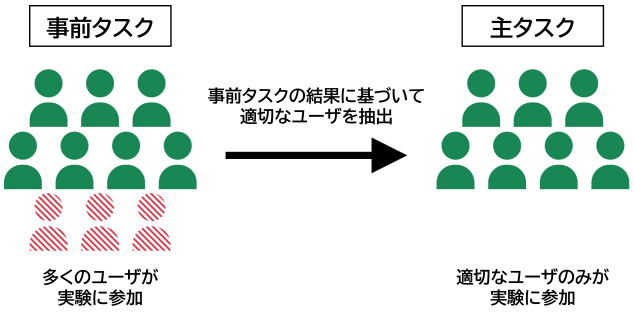

図 1 提案アプローチのイメージ図

3.2 参加者スクリーニングのための事前タスク

本研究では，主タスクとしてポインティングタスクを扱うため，その操作に関連する事前タスクを設定することで，適切な参加者をスクリーニングできると考えられる．ここで我々は，Li ら [19] が提案した物理カードと画面上のカード画像の大きさを一致させるタスク（サイズ調整タスク）に着目した．ここでは，図 2 のように参加者が画面上にサイズが標準化された物理カード（クレジットカードなど）を設置し，その大きさと一致するようにカード画像の大きさを調整する．そのため，サイズ調整の結果と物理カードの大きさの誤差をもとに操作の正確性を分析できる．

またサイズ調整タスクは，本来は視覚刺激の物理サイズを統制するために実験前に実施されるが，我々が以前に実施したサイズ調整タスクで不適切な操作や雑な操作を行う参加者が存在した．そのため，サイズ調整タスクを事前タスクとすることでそのような参加者を特定し，適切なユーザを抽出できると考えた．

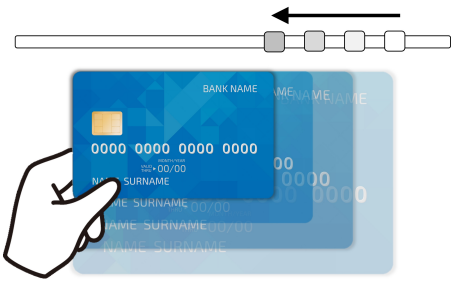

図 2 Li ら [19] によるサイズ調整タスクのイメージ図

4. 実験

4.1 実験設計

4.1.1 概要

これまでの研究と同様に，事前タスクとしてサイズ調整タスクを行い，その後に主タスクとしてポインティングタスクを行うクラウドソーシング実験を実施する．本実験では，まず実験システムにアクセスすると実験の説明を行い，その後にサイズ調整タスクを 1 回，ポインティングタスクを 4 セット行う．なお本稿では，本実験で得られたデータを使用して，サイズ調整タスクで不適切とされた参加者が混入した場合のポインティングタスクの結果についてシ

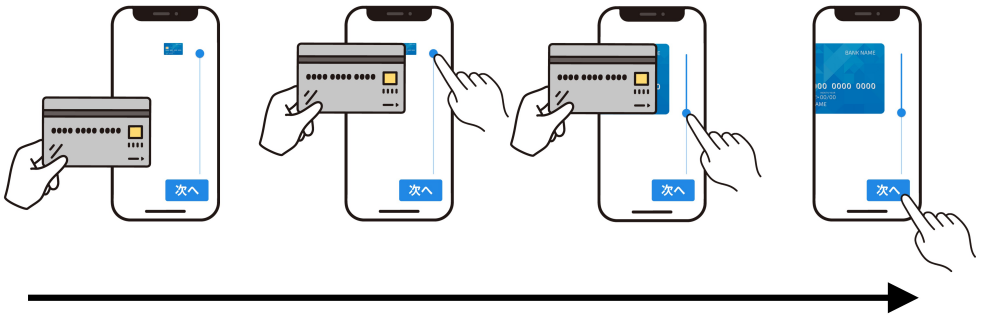

図 3 サイズ調整タスク

図 4 ポインティングタスク

ミュレーションを行い，アプローチの有用性を検証する．そのため，本実験ではすべての参加者にサイズ調整タスクとポインティングタスクの両方を行ってもらう．

また，本実験では参加者のデバイスを iPhone に限定することで，ディスプレイの画素密度を特定し，視覚刺激の表示を mm 単位で統制する．ここでは，iPhone の画面解像度データ[*1]を使用して，実験システムが取得した画面解像度をもとに参加者のデバイスの PPI（pixels per inch）を特定し，それを利用して px から mm への変換を可能としている．これにより，サイズ調整タスクの結果と物理カードの大きさの誤差を mm 単位で分析でき，ポインティングタスクにおけるターゲットのサイズも mm 単位で制御できるため，両タスクにおいて厳密な評価を行うことができる．

4.1.2 サイズ調整タスク

本実験では，図 3 のように物理カードを iPhone 上に設置し，スライダー操作によって画像をリサイズして，物理カードの短辺とカード画像の短辺の大きさを合わせる操作を行ってもらう．ここでは，クレジットカードと同じサイズ（ISO/IEC 7810 の ID-1 規格：縦 53.98 mm × 横 85.60 mm）の物理カードを使用可能とした．なお，本稿ではサイズ調整タスクをあくまでもスクリーニングのための事前タスクとして扱うため，サイズ調整タスクによって得られた画素密度は利用せず，端末情報（画面解像度）によって特定した PPI を利用して視覚刺激の大きさを統制する．

4.1.3 ポインティングタスク

図 4 のように，上下 2 か所に表示された 2 つの長方形を交互にタップするタスクを課した．2 つの長方形の表示位置（中心座標）は固定されており，現在狙うべきターゲットは水色，他方は灰色で表示される．ターゲットは横幅が画面サイズ，縦幅が W mm である．ターゲットをタップすると，上下の長方形の色が入れ替わり，新しいターゲットの縦幅が実験条件にしたがって変更される．ここでは，ターゲット外をタップした場合は成功するまで再試行させ，選択できたら次の試行に進むようにした．

本実験では，ターゲット中心間の距離 A は 30.0 mm の 1 条件，ターゲットの縦幅 W は 2.0, 2.8, 3.6, 4.4, 5.2, 6.0, 6.8, 7.6, 8.4 mm の 9 条件とした．また，「できるだけ速く」と「できるだけ正確に」の 2 つの教示を用意し，それに従って操作するように指示した [20, 21]．なお，参加者には図 4 のように非利き手でスマートフォンを保持し，利き手の人差し指で画面をタップするように指示している [14, 15]．

ポインティングタスクの 1 セットは 90 試行であり，90 試行のターゲットのうち W の 9 条件が 10 試行ずつ順番はランダムで選出される．また，合計 4 セットのうち，2 つの教示がランダムな順序で 2 回ずつ設定されるようにした．以上により，参加者 1 人あたりの総試行回数は $9W \times 10$ 試行 $\times 2$ 教示 $\times 2$ セット $= 360$ 試行 であり，1 つの（教示 $\times W$）条件について，10 試行 $\times 2$ セット $= 20$ 試行分のデータが記録される．

4.2 実験参加者

実験は Yahoo!クラウドソーシング[*2]を通じて iPhone 7 以降の機種を対象として参加者を募集し，実験を完了した参加者には 250 円の報酬を支払った．実験を完了した 584 人のうち，実験データに欠損が確認された 8 人，iPhone の PPI が特定できなかった 42 人を除外し，534 人（男性 297 人，女性 235 人，その他 2 人）が分析対象となった．

4.3 シミュレーション

4.3.1 分析概要

事前タスクによるスクリーニングの結果として，不適切とされた参加者が混入した場合の実験結果についてシミュレーションを行い，スクリーニング手法の有用性を検証する．ここでは，実験の目的が既存の GUI 操作のパフォーマンスモデルの適合度を検証することと想定し，スクリーニングによってモデル適合度が向上するかどうかを検証する．そのために，事前タスクで特定された不適切な参加者の割合を変化させたときに，ポインティングタスクのモデル適合度を算出する．

本実験で扱うモデルを式 (1) から式 (3) に示す．式中の a–d は実験で決まる回帰係数である．式 (1) はフィッツの法則であり，ターゲットまでの距離 A とターゲットの幅

[*1] https://www.ios-resolution.com/

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

W から操作時間 MT を予測するモデルである [11]．式 (2) は，ターゲットサイズに対してタップ座標の y 座標がどの程度だけ分散するかを推定するモデルであり，各 W 条件についてターゲットの中心を原点とするタップ座標の y 座標の標準偏差 σ_y を求めて回帰分析する．式 (2) で特定の W に対する σ_y を推定することで，式 (3) によってタップ座標の y 座標が幅 W のターゲット領域に入る成功率 $P(-W/2 \leq Y \leq W/2)$ を予測できる [15]．

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

$$\sigma_y^2 = c + dW^2 \tag{2}$$

$$P\left(-\frac{W}{2} \leq Y \leq \frac{W}{2}\right) = \mathrm{erf}\left(\frac{W}{2\sqrt{2}\sigma_y}\right) \tag{3}$$

これらのモデルが高い適合度を示すことは先行研究で確かめられており [11,22]，データ品質が十分に高ければ我々も同様の結果を得られるはずである．逆にデータ品質が低く，仮に参加者が W 条件に応じて MT や ER を変えないようであれば，モデル適合度は低くなると考えられる．

4.3.2 前処理

分析前に外れ値の除外を行った．まず，1 人あたりの各（教示 $\times$ W）条件における 20 試行のなかで，タップ座標が外れ値となる試行を除外した．ここでは，タップ座標がターゲット中心よりも下側ならばプラス，上側ならばマイナスとして y 座標を記録し，y 座標の標準偏差 σ を使用して，平均より 3σ 以上短いまたは 3σ 以上長い試行を除外対象とした．また，各（教示 $\times$ W）条件における 20 試行の MT についても，同様に 3σ 基準で外れ値の試行を除外した．さらに，各参加者の全 360 試行の平均 MT について同様に 3σ を基準として，外れ値となる参加者の試行をすべて除外した．以上の処理により，タップ座標について 1,042 試行，MT について 2,547 試行と 4 人の参加者が検出され，除外後の 187,347 試行（530 人）を分析対象とした．

4.3.3 分析手順

事前タスクの結果によって参加者を「合格群」と「不合格群」に分け，適切なユーザ群である合格群に主タスクの実験を依頼するのが我々のスクリーニングの目的である．その効果を検証するため，サイズ調整タスクの誤差が閾値 T mm より小さい参加者を合格群，それ以外の参加者を不合格群とする．シミュレーションでは，参加者数が合計 N の状況において不合格群の割合 X% を変化させる検証を行うため，不合格群から $N \times X$% 人，合格群から $N \times (100 - X)$% 人をランダムに抽出する．次に，抽出した参加者群のポインティングタスクの結果について，式 (1) と式 (3) のモデル適合度を算出する．以上の操作を 1,000 回繰り返してモデル適合度の平均をとる．

本シミュレーションでは N，T，X を変化させて，不適切な参加者の混入が実験結果（モデル適合度）に及ぼす影響を分析する．ここでは，N を 10, 80 人，T を 1 mm から 10 mm まで 1 mm ずつ区切った値，X を 0%から 100%まで 10%ずつ区切った値とした．閾値の変化による合格群と不合格群の人数を表 1 に示す．

表 1　閾値の変化による合格群と不合格群の人数

閾値 T (mm)	1	2	3	4	5	6	7	8	9	10
合格群 (人)	250	316	335	351	353	360	367	375	379	388
不合格群 (人)	280	214	195	179	177	170	163	155	151	142

4.4 結果と考察

4.4.1 操作時間 MT について

図 5 は，閾値 T と不合格群の割合 X を変化させ，各条件におけるポインティングタスクの MT の結果について，式 (1) に対する適合度 R^2 を評価した結果である．ここでは，ポインティングタスクの教示およびシミュレーションの参加者数 N ごとに結果をヒートマップで示している．図 5 において，各ヒートマップの左上から右下にかけてモデル適合度が低下しており，閾値を緩く設定した場合に不合格群の割合が多くなると，モデル適合度が低下する傾向がみられる．ここで，閾値を緩く設定した場合は，サイズ調整の誤差が大きく操作が雑な参加者が不合格群となり，そのような参加者が増加することでモデル適合度が低下していると考えられる．この結果は，これまでの研究とは異なり，不合格群の割合が MT を予測するモデルの評価にも影響を及ぼすことを示しており，不合格群の割合を少なくするほどモデル適合度が向上する傾向があることから，スクリーニング手法の有用性が示唆されたといえる．

この結果の差異は，デバイス条件の変更が大きな要因と考えられる．これまでの研究では PC でマウスを使用しての実験であり，不真面目な参加者が実験を短時間で終了させるために素早くターゲットをクリックしたとしてもターゲット間の距離に応じたポインターの移動が必要となり，比較的正常な操作時間になりやすいと考えられる．一方，本実験はスマートフォンでの実験であり，ターゲットが出現する位置も固定されていた．そのため，不真面目な参加者が実験を短時間で終了させるため，指を 2 本使ったり両手を使ったりすることで指の移動の必要がなくターゲットをタップできる．実験の指示として各参加者には図 4 のように非利き手でスマートフォンを保持し，利き手の人差し指で画面をタップするように指示したが，不真面目な参加者にはその指示が機能せず，正常でない操作時間となった可能性がある．以上より，スマートフォンのような参加者の操作方法の統制が難しい環境でのクラウドソーシング実験では，式 (1) のような MT のモデルを検証する目的において，スクリーニング手法の有用性が高くなるといえる．

また，図 5 において，参加者数 $N = 80$ では $N = 10$ と比較して全体的にモデル適合度が高く，不合格群の割合が多い場合のモデル適合度も向上しており，参加者数の増加によって不合格群による影響が抑制されているといえる．

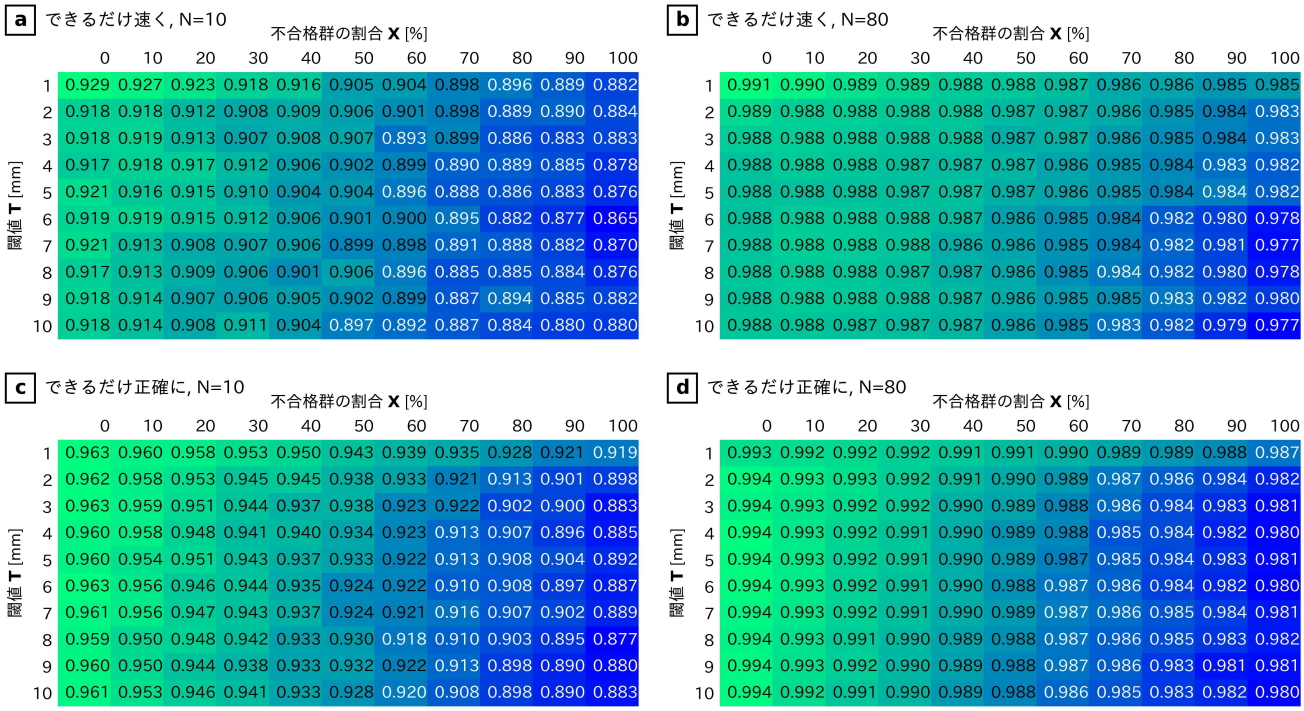

図 5 MT について式 (1) に対するモデル適合度

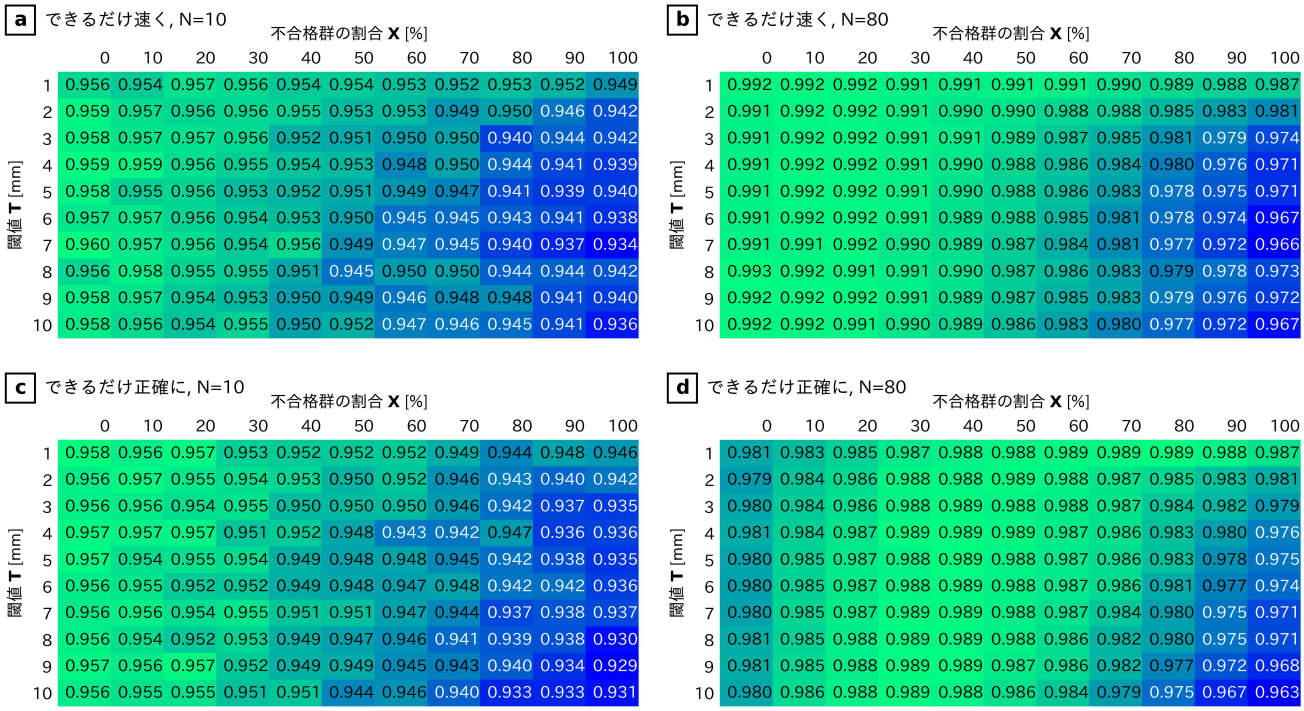

図 6 ER について式 (3) に対するモデル適合度

しかし，不合格群の割合によってモデル適合度は変動するため，本研究におけるスクリーニング手法は，不適切な参加者の混入による影響を抑えるには参加者数が十分でない場合に，モデル適合度の向上に寄与できると考えられる．

4.4.2 エラー率 ER について

図 6 は，ER のシミュレーション結果で，式 (3) に対する適合度 R^2 を評価した結果を示したものである．図 6 において，各ヒートマップの左上から右下にかけてモデル適合度が低下しており，閾値を緩く設定した場合に不合格群の割合が多くなると，モデル適合度が低下する傾向がみられる．そのため，これまでの研究で観測された適合度の低下が本実験でも観測されたことになる．本実験ではターゲット幅の条件や 1 人あたりの繰り返し回数を増やして厳密な検証を行ったため，スクリーニング手法の有用性がより強く裏付けられたといえる．なお，参加者数 N による影響は前項で述べた傾向と同様の傾向がみられている．

4.5 制約

本実験ではデバイスを iPhone に限定することで，サイズ調整タスクの結果と物理カードの大きさの誤差を mm 単位で分析し，ポインティングタスクにおけるターゲットの幅も mm 単位で統制して，両タスクで厳密な検証を行うことができた．しかし，図 5 や図 6 にみられるように，とくに ER モデルの場合に各ヒートマップにおいて最も左上に位置する閾値 $T = 1$ mm，不合格群の割合 $X = 0\%$ の領域でモデル適合度が最良ではない場合が存在した．この領域は，$T = 1$ が最も厳しい閾値であり，$X = 0$ では不合格群に属する参加者は存在しないため，実験実施者にとって理想的な参加者群だと考えられる．そのような場合にモデル適合度が最良とならない要因として，本実験ではポインティングタスクでエラーが発生した場合に成功するまで再試行させる処理を行っていたことが挙げられる．このような処理方法では，不真面目な参加者が短時間で実験を終了させるためには速くかつ正確にタップする戦略をとり，真面目な参加者と操作の区別がつきにくいと考えられる．

5. 追加実験

5.1 実験設計

前章の実験を踏まえて，追加実験ではポインティングタスクでエラーが発生した場合でもすぐに次の試行に進む処理を採用する．これにより，前章において不真面目な参加者が短時間で実験を終了させるためには速くかつ正確にタップする戦略をとり，真面目な参加者と操作の区別がつきにくいという制約を解消することを目指す．追加実験は，エラーの処理方法を除いて前章と同様の実験設計である．

5.2 実験参加者

575 人が実験を完了した．このうち，実験データに欠損が確認された 1 人，iPhone の PPI が特定できなかった 55 人を除外し，519 人（男性 250 人，女性 266 人，その他 3 人）が分析対象となった．

5.3 シミュレーション

前章と同様のシミュレーションを行う．ポインティングタスクの外れ値として，タップ座標について 1,011 試行，MT について 2,717 試行と 4 人の参加者が検出され，除外後の 181,796 試行（515 人）を分析対象とした．閾値 T の変化による合格群と不合格群の人数を表 2 に示す．

表 2 閾値の変化による合格群と不合格群の人数

閾値 T (mm)	1	2	3	4	5	6	7	8	9	10
合格群 (人)	249	307	327	336	351	358	359	362	368	373
不合格群 (人)	266	208	188	179	164	157	156	153	147	142

5.4 シミュレーション結果

5.4.1 操作時間 MT について

図 7 は，追加実験における MT のシミュレーション結果である．ここでは，各ヒートマップの左上から右下にかけてモデル適合度が低下しており，閾値を緩く設定した場合に不合格群の割合が多くなると，モデル適合度が低下する傾向がみられる．そのため，前章で観測された適合度の低下が追加実験でも観測されたことになる．

前章と比較すると，追加実験では不合格群の割合の増加による適合度の低下幅がやや大きいことが差異であった．具体的には，前章の図 5 では $N = 10$ の場合でモデル適合度が低下してもほとんどが 0.88 を超える値となっていたが，追加実験における図 7 では $N = 10$ の場合で 0.65 付近まで低下する場合もある．このような差異は，エラーの場合でもすぐに次の試行に進む処理によるものであり，真面目な参加者と不真面目な参加者の操作の区別がつきにくいという制約が解消され，モデル適合度の変化が顕著になったと考えられる．そのため，追加実験のように参加者が指示を守らなくてもタスクを完了できる状況では，スクリーニング手法の有効性が高まるといえる．

5.4.2 エラー率 ER について

図 8 は，追加実験における ER のシミュレーション結果である．ここでは，各ヒートマップの左上から右下にかけてモデル適合度が低下しており，閾値を緩く設定した場合に不合格群の割合が多くなると，モデル適合度が低下する傾向がみられる．そのため，これまでの研究や前章で観測された適合度の低下が追加実験でも観測されたことになる．

前章との差異として，追加実験では不合格群の割合の増加による適合度の低下幅がやや大きいことが挙げられる．前章の図 6 ではモデル適合度が低下してもほとんどが 0.93 を超える値であったが，追加実験における図 8 では $N = 10$ の場合で 0.85 付近まで低下する場合もある．また，図 8 では図 6 と比較して，不合格群の割合が少ない場合でも適合度の低下が発生する傾向がみられる．さらに，各ヒートマップにおいて最も左上に位置する閾値 $T = 1$，不合格群の割合 $X = 0$ の領域でモデル適合度が最良となることも差異として挙げられる．前章の図 6 では，各ヒートマップにおける $T = 1$，$X = 0$ の領域でモデル適合度が最良ではない場合があったが，図 8 においてはほとんどの場合で最良となり，T と X による適合度の変化が顕著になった．

このような差異は，エラーの場合でもすぐに次の試行に進む処理によるもので，不真面目な参加者はエラーを多くすることを受け入れて短時間に操作し，ER モデルが想定しているポインティング操作から逸脱した可能性がある．そのため，T と X によるモデル適合度の変化が顕著になり，各ヒートマップにおいて最も左上に位置する閾値 $T = 1$，不合格群の割合 $X = 0$ の場合にモデル適合度が最良となったと考えられる．

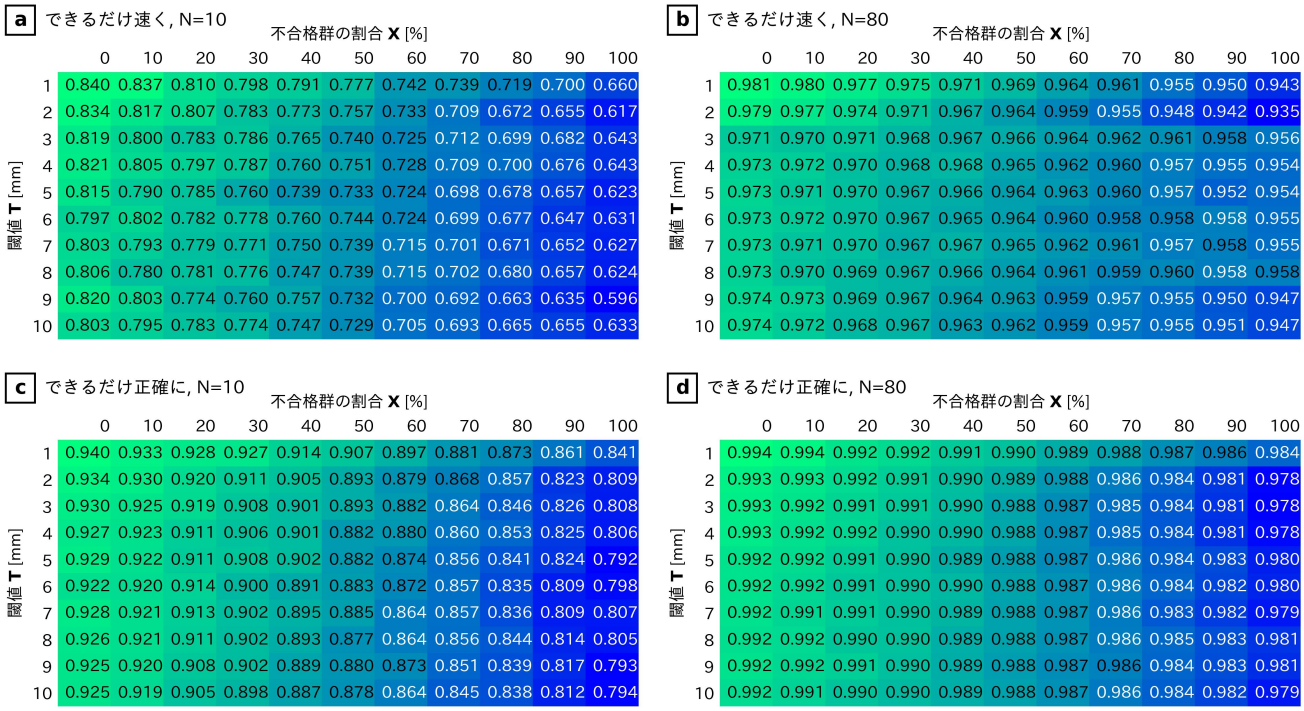

図 7 MT について式 (1) に対するモデル適合度

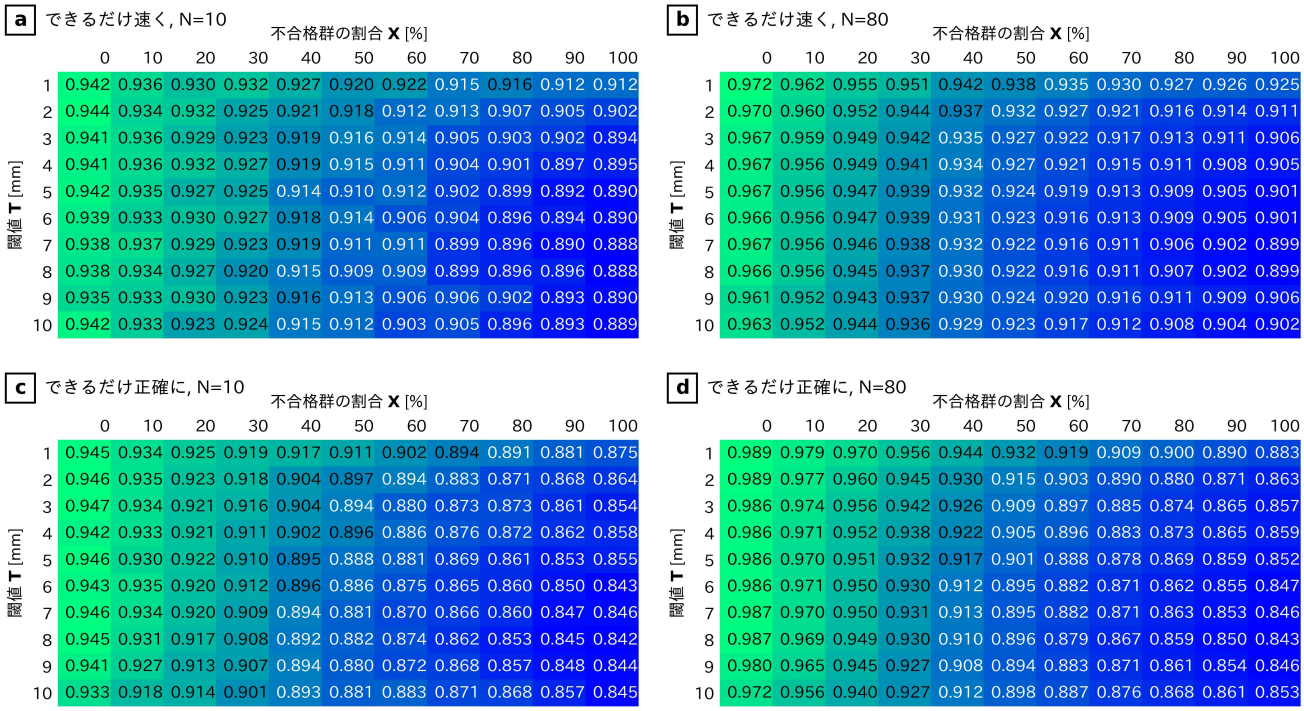

図 8 ER について式 (3) に対するモデル適合度

6. 課題と今後の展望

本稿では，サイズ調整タスクを利用したスクリーニング手法について，スマートフォン環境で視覚刺激の表示を mm 単位で統制した実験を通じて厳密な評価を行った．その結果，これまでの研究と同様にスクリーニング手法は GUI 操作モデルの適合度を向上させることを再確認し，追加実験によって手法がより有効に機能する場面が明らかとなり，その有用性がより明確に示された．一方で，スクリーニング手法には課題もあり，さらなる検証が求められる．

本研究のスクリーニングは，事前タスクでは適切に操作したが主タスクでは不適切に操作した参加者を特定できないという制約をもつ．そのため，スクリーニング後に構成される理想的と見なされる参加者群にも，主タスクの観点ではノイズとなる参加者が混入する可能性がある．また，サイズ調整タスクを利用したスクリーニングでは閾値 T の

設定が不可避であるが，T や不合格群の割合 X を変化させるとモデル適合度が段階的に変化するため，適切な T を一意に決定することは困難である．T を厳しく設定すれば不適切な参加者の混入をより抑制できる一方で，参加者数が減少するトレードオフが生じてしまう．また，本稿の実験結果や考察は，事前タスクをサイズ調整，主タスクをポインティングにした場合に限定され，一方または両方を他のタスクにしても同様にスクリーニングが有効であることを確かめるにはさらなる調査が必要である．

本稿の 2 つの実験では，不合格群の割合を低くすることでモデル適合度が向上する傾向が一貫して観察され，不適切な参加者を判定するための情報として主タスクの結果を用いずに，事前タスクのみの UI 操作の結果によってスクリーニングができた．この結果は事前タスクが主タスクと異なるタスクの場合も適切にスクリーニングできることを示唆しており，主タスクを変更するたびに事前タスクを変更する必要はなく，ステアリングの法則など他の GUI 実験にもスクリーニングが適用できると考えられる．今後はサイズ調整タスクを利用したスクリーニング手法が他の GUI 実験にも適用可能かどうかの検証も行う予定である．

参考文献

[1] Yamanaka, S.: Utility of crowdsourced user experiments for measuring the central tendency of user performance to evaluate error-rate models on guis, *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9, pp. 155–165 (2021).

[2] 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*, pp. 6813–6824 (2017).

[3] 三山貴也，中村聡史，山中祥太：Web ベースの実験における事前タスクを用いたユーザ分類の検討，情報処理学会 研究報告 HCI，Vol. 2025-HCI-211, No. 14, pp. 1–8 (2025).

[4] 三山貴也，中村聡史，山中祥太：クラウドソーシングを活用した GUI 実験における参加者スクリーニング手法の検証，情報処理学会 研究報告 HCI，Vol. 2025-HCI-214, No. 14, pp. 1–8 (2025).

[5] Brühlmann, F., Petralito, S., Aeschbach, L. F. and Opwis, K.: The quality of data collected online: An investigation of careless responding in a crowdsourced sample, *Methods in Psychology*, Vol. 2, p. 100022 (2020).

[6] Curran, P. G.: Methods for the detection of carelessly invalid responses in survey data, *Journal of Experimental Social Psychology*, Vol. 66, pp. 4–19 (2016).

[7] Oppenheimer, D. M., Meyvis, T. and Davidenko, N.: Instructional manipulation checks: Detecting satisficing to increase statistical power, *Journal of experimental social psychology*, Vol. 45, No. 4, pp. 867–872 (2009).

[8] Komarov, S., Reinecke, K. and Gajos, K. Z.: Crowdsourcing performance evaluations of user interfaces, *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 207–216 (2013).

[9] Grossman, T. and Balakrishnan, R.: The bubble cursor: enhancing target acquisition by dynamic resizing of the cursor's activation area, *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 281–290 (2005).

[10] 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, p. 381 (1954).

[11] Soukoreff, R. W. and 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*, Vol. 61, No. 6, pp. 751–789 (2004).

[12] Meyer, D. E., Abrams, R. A., Kornblum, S., Wright, C. E. and Keith Smith, J.: Optimality in human motor performance: ideal control of rapid aimed movements., *Psychological review*, Vol. 95, No. 3, p. 340 (1988).

[13] Wobbrock, J. O., Cutrell, E., Harada, S. and MacKenzie, I. S.: An error model for pointing based on Fitts' law, *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1613–1622 (2008).

[14] Bi, X. and Zhai, S.: Predicting finger-touch accuracy based on the dual Gaussian distribution model, *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 313–319 (2016).

[15] Yamanaka, S. and Usuba, H.: Rethinking the dual gaussian distribution model for predicting touch accuracy in on-screen-start pointing tasks, *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4, No. ISS, pp. 1–20 (2020).

[16] 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, *Proceedings of the 22nd international ACM Sigaccess conference on computers and accessibility*, pp. 1–15 (2020).

[17] Yamanaka, S.: Test-Retest Reliability on Movement Times and Error Rates in Target Pointing, *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pp. 178–188 (2022).

[18] Yamanaka, S., Kinoshita, T., Oba, Y., Tomihari, R. and Miyashita, H.: Varying subjective speed-accuracy biases to evaluate the generalizability of experimental conclusions on pointing-facilitation techniques, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–13 (2023).

[19] Li, Q., Joo, S. J., Yeatman, J. D. and Reinecke, K.: Controlling for participants' viewing distance in large-scale, psychophysical online experiments using a virtual chinrest, *Scientific reports*, Vol. 10, No. 1, p. 904 (2020).

[20] Zhai, S., Kong, J. and Ren, X.: Speed–accuracy tradeoff in Fitts' law tasks—on the equivalency of actual and nominal pointing precision, *International journal of human-computer studies*, Vol. 61, No. 6, pp. 823–856 (2004).

[21] Yamanaka, S.: Relative merits of nominal and effective indexes of difficulty of Fitts' law: Effects of sample size and the number of repetitions on model fit, *International Journal of Human–Computer Interaction*, Vol. 41, No. 1, pp. 574–591 (2025).

[22] Yamanaka, S. and Usuba, H.: Tuning Endpoint-variability Parameters by Observed Error Rates to Obtain Better Prediction Accuracy of Pointing Misses, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18 (2023).