Vol.2026-HCI-220 No.34
Vol.2026-UBI-92 No.34
2026/11/27

GUI操作を伴うオンライン実験における参加者スクリーニングのための評価手法の比較

三山 貴也[1]　中村 聡史[1]　山中 祥太[2]

概要：GUI 操作を伴うオンライン実験では，クラウドソーシングによって多くの参加者を素早く募集でき，GUI 操作時のデータを大量に収集できる．一方，指示を守らない参加者や雑な操作を行う参加者が存在するため，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，参加者をスクリーニングするため，実験前の GUI 操作タスクの結果を評価指標とする手法を提案し，提案手法を用いたスクリーニングによって 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，これまでの研究は提案手法の検証にとどまり，文章中の特殊な指示を実行したかを評価指標とする既存手法との比較ができていなかった．そこで本稿では，提案手法と既存手法について，スクリーニングによるモデル適合度の変化と参加者による各手法への印象評価を比較した．その結果，いずれの手法でもスクリーニングによるモデル適合度の向上がみられ，提案手法は既存手法よりもわずらわしさが少ないと評価された．また，各手法でスクリーニングの対象となった参加者の一致度は低く，評価の観点が異なる可能性が示唆された．

1. はじめに

グラフィカルユーザインタフェース（GUI）の操作を伴うユーザ実験において，クラウドソーシングを利用したオンライン実験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，GUI 操作を伴うオンライン実験において，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] において，GUI 操作を伴うオンライン実験で参加者をスクリーニングするため，実験前の事前タスクの結果を評価指標とする手法を提案してきた．ここでは，事前タスクとして画像をリサイズするタスク，実験のメインタスクとしてポインティングタスクを実施した結果，提案手法を用いたスクリーニングによってメインタスクで 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，これまでの研究は提案手法の検証にとどまり，既存の評価手法との比較ができていなかった．

そこで本稿では，提案手法と既存手法を比較し，提案手法の特徴を明らかにすることを目指す．ここで，オンライン実験におけるスクリーニングでは，文章中の特殊な指示を実行したかを評価指標とする手法が用いられている．本稿では，その代表的な手法である IMC（Instructional Manipulation Check）[4] および IRI（Instructed Response Item）[5] を比較対象とし，各手法を用いたスクリーニングに伴うメインタスクのモデル適合度の変化を比較する．また，既存手法はタスクへの注意を促すと評価される一方でわずらわしさなどの点で否定的に評価されることもある [6] ため，各手法に対する参加者の印象を比較することで参加者の視点から提案手法の特徴を整理する．さらに，各手法でスクリーニングの対象となる参加者が一致するかを分析し，各手法による評価の違いについて検討を行う．

2. 関連研究

2.1 オンライン実験のデータ品質

オンライン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 [7] 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 [8]．また Oppenheimer ら [4] は，実験の説明文に埋め込まれた特殊な指示を参加

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

© 2026 Information Processing Society of Japan

者が実行したかチェックする 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実験結果が異なる可能性があるとしている．Gummer ら [5] は，アンケートの設問群の中に特殊な指示の設問を埋め込む IRI を検証し，IRI を通過しなかった参加者群では無回答や回答の不整合が多いとしている．

以上のように，オンライン実験のデータ品質に関する検証が行われている．本研究では，実験前の GUI 操作タスクの結果を評価指標として参加者をスクリーニングすることで，これまでの取り組みをさらに改善する可能性を探る．

2.2 GUI 操作を伴うオンライン実験

GUI 操作を伴う実験について，オンライン実験と実験室実験の結果の比較が行われている．Komarov ら [9] は，Bubble Cursor [10] 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，オンライン実験でも再現されたと報告している．また Findlater ら [2] は，マウスとタッチ操作のポインティングタスクでオンライン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，GUI 操作を伴うオンライン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．本研究の提案手法は，GUI に関する実験特有のインタラクションに適したタスクをスクリーニングに利用するものであり，実験の特性を踏まえたアプローチとして位置づけられる．

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，クリックまでの時間 MT を予測できる [11]．またエラー率 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 [12, 13]．

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．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 [14]．

3. 提案手法

3.1 概要

本研究では，GUI 操作を伴うオンライン実験において，実験前の事前タスクによって参加者をスクリーニングし，本来目的とする実験（メインタスク）を事前タスクの基準を満たした参加者群のみで実施するアプローチをとる．具体的には，メイン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をスクリーニングの評価指標として用いることで，メインタスクで適切な操作を行う参加者の割合が高まることが期待される．

3.2 事前タスク

本研究では，メイン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，適切にスクリーニングができると考えられる．ここで我々は，Li ら [15] 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．具体的には，図 1 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，我々が以前に実施したサイズ調整タスクにおいて誤差が大きく適切に操作をしていないと思われる参加者が存在した．以上のことから，サイズ調整タスクにおける誤差をスクリーニングの評価指標として用いることを考えた．

4. 実験 1：IMC との比較

4.1 実験設計

4.1.1 概要

実験 1 では，サイズ調整タスク（提案手法）と IMC（既存手法）の比較を行う．これまでの研究 [3] では，事前タスクをサイズ調整タスク，メインタスクをポインティングタスクとしたオンライン実験を実施し，収集したデータを使用したシミュレーションによって提案手法を用いたスクリーニングについて検証してきた．本実験では，これまでの研究の手続きに加えて事前タスクとして IMC を追加し，1 人の参加者がサイズ調整タスクと IMC の両方を実施することで，両手法について参加者内で比較を行う．

本実験は，2 種類の事前タスク（サイズ調整タスク，IMC），メインタスク（ポインティングタスク），事後アンケートの順に実施される．事前タスクについては，サイズ調整タスクを先に行う群と IMC を先に行う群を設け，各群には参加者のほぼ半数が含まれるようにする．事後アンケートでは，2 種類の事前タスクのそれぞれに対する印象評価を行ってもらう．なお，これまでの研究と同様にシミュレーションによって検証を行うため，全参加者にすべてのタスクを実施してもらいデータを収集する．また，実験は iPhone のみを対象として実施し，iPhone のディスプレイ情報を利用して視覚刺激を mm 単位で統制している．

4.1.2 サイズ調整タスク

図 1 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

© 2026 Information Processing Society of Japan

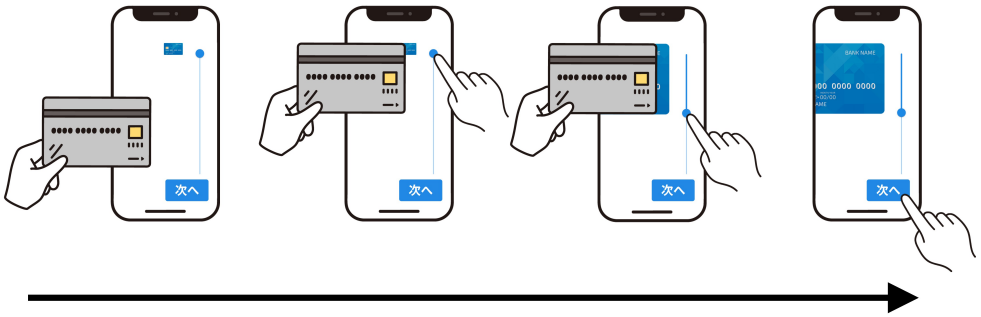

図 1 サイズ調整タスク



図 2 ポインティングタスク

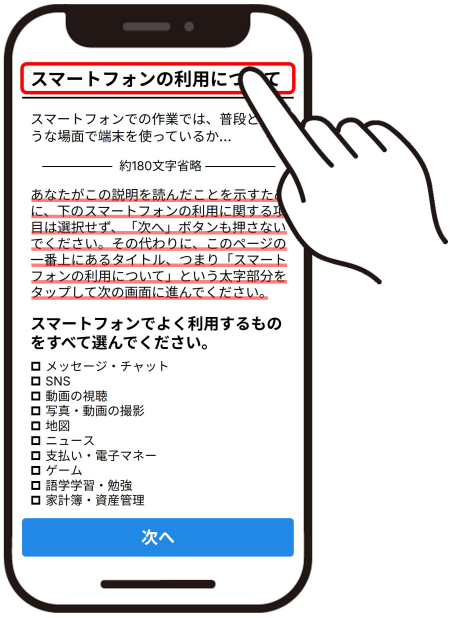

図 3 IMC

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によって得られた画素密度は利用せず，iPhone のディスプレイ情報を利用して視覚刺激の大きさを統制する．

4.1.3 IMC

Oppenheimer ら [4] と同様に，実験の説明文に埋め込まれた特殊な指示を参加者が実行したかチェックする．図 3 のように，画面には約 350 文字の説明文，アンケート設問，および「次へ」ボタンを提示した．ここでは，通常はアンケート設問において回答を選択して「次へ」ボタンを押す構成だが，説明文の後半では回答をせず「次へ」ボタンも押さずにページ上部のタイトルをタップするよう指示しており，参加者は通常とは異なる操作を行う必要がある．

4.1.4 ポインティングタスク

図 2 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．なお，エラー時には再試行をさせず，すぐに次の試行に進むようにした．

表 1 事後アンケート

番号	内容
Q1	この項目の内容を理解できた．
Q2	この項目は，わかりにくいと感じた．
Q3	この項目は，わずらわしいと感じた．
Q4	この項目によって，その後の説明文をより注意して読み，気をつけて操作しようと思った．
Q5	この項目によって，その後の回答や操作に取り組むモチベーションが上がった．
Q6	この項目によって，自分が管理されているように感じた．
Q7	この項目によって，自分の回答や操作を誘導されているように感じた．
Q8	この項目のように，回答や操作の方法を指示されたくないと感じた．

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 [14,16]．なお，参加者には図 2 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 [12,13]．

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．

4.1.5 事後アンケート

2 種類の事前タスクへの印象を表 1 の 8 問で評価してもらう．評価に用いる設問は，Silber ら [6] による IRI への評価を参考にした．各設問は 1「まったくそう思わない」から 5「とてもそう思う」の 5 段階で回答するものであり，Q1 のみ「この項目がそのような内容だったことに気づかなかった」という第 6 の選択肢を設けた．

4.2 実験参加者

実験は Yahoo!クラウドソーシング[*1]を通じて iPhone X 以降の機種を対象に参加者を募集し，実験を完了した参加

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

© 2026 Information Processing Society of Japan

者には 300 円の報酬を支払った．実験を完了した 489 人のうち，iPhone のディスプレイ情報が特定できなかった 16 人を除外し，473 人（男性 249 人，女性 220 人，その他 4 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

サイズ調整タスクと IMC それぞれを用いたスクリーニングを評価するため，各手法で基準を満たさなかっ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ミュレーションを行う．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．評価手法の基準を満たさなかった参加者の割合が高いほどモデル適合度が低下するのであれば，その手法を用いたスクリーニングによってモデル適合度の向上が期待できると考えられる．

本実験で扱うモデルを式 (1) から式 (2) 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 W から操作時間 MT を予測するモデルである [11]．式 (2) は，タップ座標の y 座標が幅 W のターゲット領域に入らない確率（エラー率）ER を予測するモデルである [13]．ここでは，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 を推定することでエラー率を算出する．

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

$$ER = 1 - \mathrm{erf}\left(\frac{W}{2\sqrt{2}\sigma_y}\right),\ \sigma_y^2 = c + dW^2 \tag{2}$$

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 [11,17]，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 ポインティングタスクの前処理

これまでの研究 [3] と同様にポインティングタスクの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，平均から 3σ 以上離れた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．以上により，試行単位で 3,185 試行，参加者単位で 0 人が検出され，除外後の 167,095 試行（473 人）を分析対象とした．

4.3.3 分析手順

本実験では，事前タスクによって参加者をスクリーニングし，その基準を満たした参加者群でメインタスクを実施することを想定している．そこでスクリーニングの効果を検証するため，各手法の基準を満たした参加者を「通過群」，それ以外の参加者を「不通過群」として，ポインティングタスクの分析対象に占める不通過群の割合を変化させたときのモデル適合度を比較する．サイズ調整タスクでは誤差が閾値 T mm より小さい場合は通過群，それ以外は不通過群とし，IMC では説明文中の特殊な指示を実行した場合は通過群，それ以外は不通過群とする．

シミュレーションでは，参加者数が合計 N の状況において不通過群の割合 X% を変化させる検証を行うため，不通過群から $N \times X$% 人，通過群から $N \times (100 - X)$% 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 (2) 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．ここでは，N を 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%ずつ区切った値で変化させた．

4.4 結果と考察

サイズ調整タスクの閾値ごとの通過群と不通過群の人数を表 2 に示す．IMC の通過群は 186 人（39.3%），不通過群は 287 人（60.7%）であった．

表 2 サイズ調整タスクの閾値ごとの通過群と不通過群の人数

	閾値 T (mm)									
群	1	2	3	4	5	6	7	8	9	10
通過 (人)	254	315	329	342	349	353	358	361	366	371
不通過 (人)	219	158	144	131	124	120	115	112	107	102

4.4.1 スクリーニングに伴うモデル適合度の変化

図 4 と図 5 は，閾値 T と不通過群の割合 X を変化させたときのポインティングタスクの結果について，それぞれ式 (1) と式 (2) 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条件ごとに結果をヒートマップで示しており，IMC の行は IMC の評価基準を用いて X のみを変化させた結果である．

図 4 の MT についての結果では，サイズ調整タスクではいずれの教示条件でも X が大きいほど R^2 が低下する傾向がみられ，これまでの研究と同様の傾向が確認された．一方 IMC では，$X = 0$% から 100% への変化に伴い，正確さ重視では 0.993 から 0.974 に低下したものの，速さ重視では 0.971 から 0.970 と変化が小さかった．また図 5 の ER についての結果では，両手法ともいずれの教示でも X が大きいほど R^2 が低下する傾向がみられ，MT と比較すると R^2 の低下幅が大きい傾向にあった．

以上のことから，サイズ調整タスクと IMC の両手法ともに，特に ER モデルの適合度についてスクリーニングに

© 2026 Information Processing Society of Japan

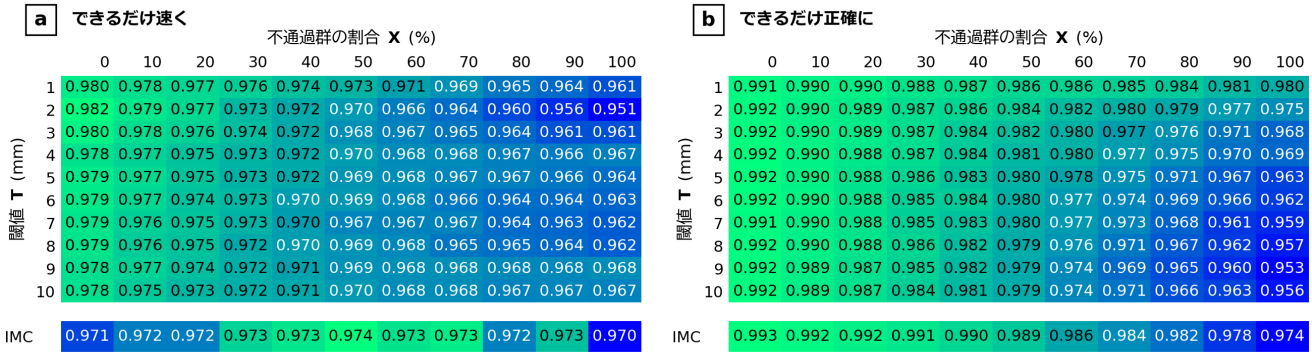

図 4 *MT* について式 (1) に対するモデル適合度

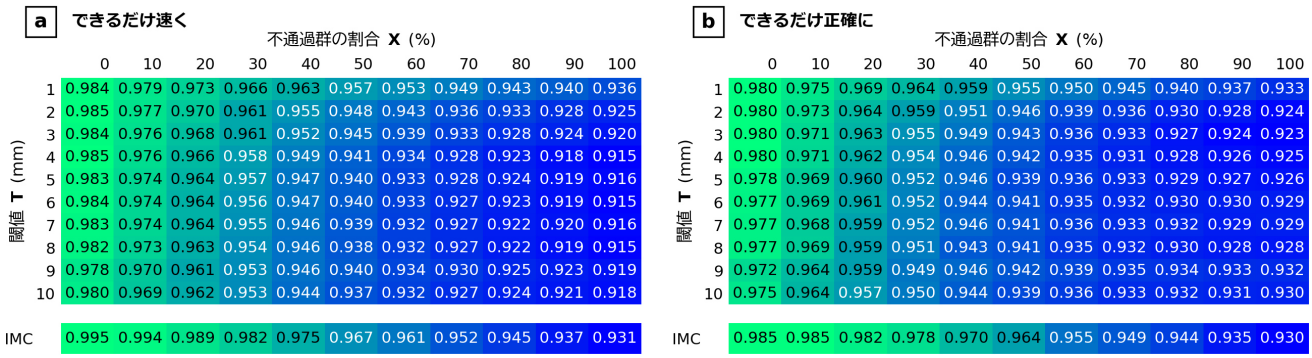

図 5 *ER* について式 (2) に対するモデル適合度

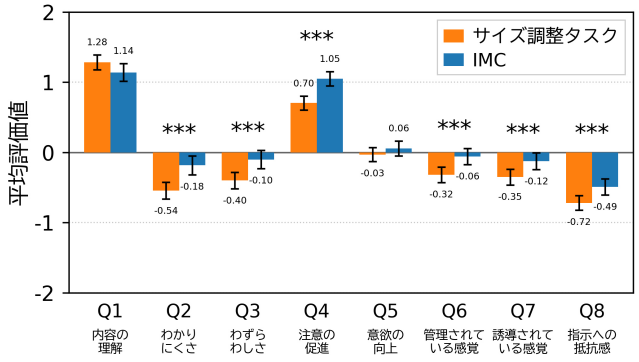

図 6 サイズ調整タスクと IMC の印象評価

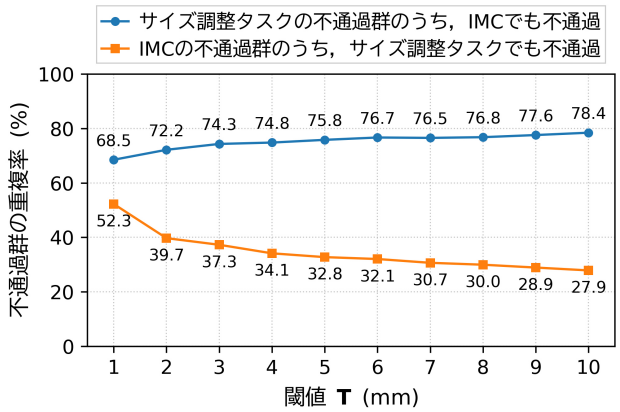

図 7 サイズ調整タスクと IMC の不通過群の重複率

よる向上が期待できると考えられる．一方で *MT* モデルについては，IMC によるスクリーニングでモデル適合度がほとんど変化しない場合もあり，スクリーニングの効果は検証するモデルや教示条件によって異なることが示唆された．提案手法では，これまでの研究と同様に，モデルや教示条件によらず不通過群の割合が高いほどモデル適合度が低下する傾向がみられたため，IMC よりもスクリーニングの効果が比較的一貫していると考えられる．

4.4.2 事前タスクの印象評価

サイズ調整タスクと IMC の印象評価の結果を図 6 に示す．なお，1 から 5 の評価値から 3 を引き，-2 から 2 の範囲に値を変換している．Q1 は両タスクで「この項目がそのような内容だったことに気づかなかった」という選択肢以外を選んだ 380 人，Q2–Q8 は全 473 人を対象に，各設問を対応のある両側 Wilcoxon 符号付順位検定で比較し，8 項目の p 値を Holm 法で補正した．

図 6 より提案手法は IMC と比較して，わかりにくさ（Q2），わずらわしさ（Q3），管理されている感覚（Q6），誘導されている感覚（Q7），指示への抵抗感（Q8）といった否定的な印象に関する項目で評価値が有意に低かった．一方で提案手法は，肯定的な印象に関する項目である注意の促進（Q4）の評価値も IMC より有意に低かった．また，内容の理解（Q1）と意欲の向上（Q5）については IMC との間に有意差がなかった．

以上のことから，提案手法は参加者が感じるわずらわしさや指示への抵抗感といったその後のタスクに影響を与え

© 2026 Information Processing Society of Japan

る可能性のある否定的な感情に配慮してスクリーニングをする際に，有用な選択肢となる可能性がある．一方，注意の促進の観点では IMC より低く評価されたことから，参加者が事前タスクに抱く感情とタスクへの注意の促進という両面を踏まえて，実験の目的に応じた評価手法を選択する必要があると考えられる．

4.4.3 スクリーニングの対象となる参加者の一致度

両手法を用いたスクリーニングによってモデル適合度の向上が期待できる一方で，不通過と判定される参加者が同じとは限らない．そこで，両手法でスクリーニングの対象となる参加者がどの程度重なるかを確認するため，サイズ調整タスクと IMC のそれぞれで不通過群とされた参加者の重複率を分析した結果を図 7 に示す．

図 7 より，サイズ調整タスクの不通過群のうち IMC でも不通過となった割合は，$T = 1, 5, 10$ mm でそれぞれ 68.5%，75.8%，78.4%であった．一方，IMC の不通過群のうちサイズ調整タスクでも不通過となった割合は，同じ閾値でそれぞれ 52.3%，32.8%，27.9%であった．また，不通過群同士の重なりを示す Jaccard 係数は，同じ閾値でそれぞれ 0.421，0.297，0.259 であった．さらに，偶然による一致を補正した判定全体の一致度である Cohen の κ は，$T = 1$–10 mm の範囲で 0.137–0.153 と低かった．

以上のことから，提案手法と IMC ではスクリーニングの対象となる参加者の一致度が低く，それぞれの評価の観点が異なる可能性がある．提案手法では GUI 操作の正確さを，IMC では教示文中の特殊な指示を実行したかを評価指標としており，こうしたタスクの内容や判定基準の違いが判定の不一致に関係した可能性がある．そのため，評価手法の選択にあたっては，その評価手法がどのような観点で評価をしているか，またその観点がメインタスクの内容と整合しているかを考慮する必要があると考えられる．

5. 実験 2：IRI との比較

実験 2 では，サイズ調整タスク（提案手法）と IRI（既存手法）の比較を行う．実験 1 の IMC を IRI に置き換え，その他の実験手続きは実験 1 と共通である．実験を完了した 495 人のうち，iPhone のディスプレイ情報が特定できなかった 19 人を除外し，さらに 4.3.2 項と同様の処理によりポインティングタスクの 3,358 試行と 2 人の参加者を除外した．その結果，474 人（男性 246 人，女性 223 人，その他 5 人），167,304 試行が分析対象となった．

5.1 IRI

Silber ら [6] を参考に，選択式のアンケート全 9 問のうち第 5 問に特殊な指示を含めた（図 8）．第 5 問では，決済アプリの利用頻度にかかわらず「利用したことがない」を選択するよう指示し，この設問で指定された選択肢を選んだ参加者が通過群，それ以外が不通過群となる．

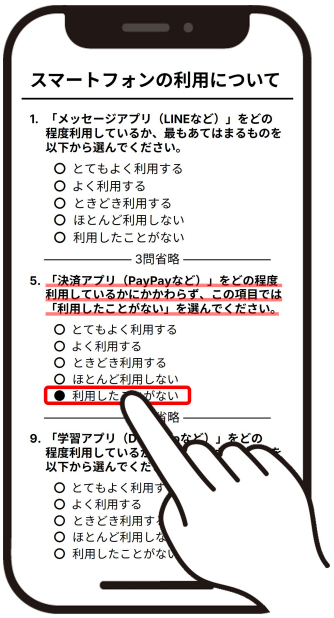

図 8 IRI

5.2 結果と考察

サイズ調整タスクの閾値ごとの通過群と不通過群の人数を表 3 に示す．IRI の通過群は 158 人（33.3%），不通過群は 316 人（66.7%）であった．

表 3 サイズ調整タスクの閾値ごとの通過群と不通過群の人数

	閾値 T (mm)									
群	1	2	3	4	5	6	7	8	9	10
通過 (人)	246	301	326	342	346	350	353	360	365	367
不通過 (人)	228	173	148	132	128	124	121	114	109	107

5.2.1 スクリーニングに伴うモデル適合度の変化

図 9 と図 10 に，MT と ER のモデル適合度 R^2 についての結果を示す．サイズ調整タスクでは，実験 1 と同様の傾向が確認された．一方，IRI の MT については，$X = 0\%$ から 100% への変化に伴い，速さ重視では R^2 が 0.975 から 0.965 に低下したものの，正確さ重視では 0.991 から 0.990 と変化が小さかった．また，ER についてはいずれの教示条件でも X が大きいほど R^2 が低下する傾向がみられ，MT と比較すると R^2 の低下幅が大きい傾向にあった．

以上のことから，IRI を用いたスクリーニングについても，特に ER モデルの適合度の向上が期待できると考えられる．一方で IRI も実験 1 の IMC と同様に，スクリーニングによってモデル適合度がほとんど変化しない場合もあり，提案手法のスクリーニングの効果が比較的一貫していることが示唆されたといえる．

5.2.2 事前タスクの印象評価

サイズ調整タスクと IRI の印象評価の結果を図 11 に示す．Q1 は両タスクで「この項目がそのような内容だったことに気づかなかった」という選択肢以外を選んだ 364 人，Q2–Q8 は全 474 人を対象とした．

図 11 より提案手法は IRI と比較して，わかりにくさ（Q2），わずらわしさ（Q3），管理されている感覚（Q6），

© 2026 Information Processing Society of Japan

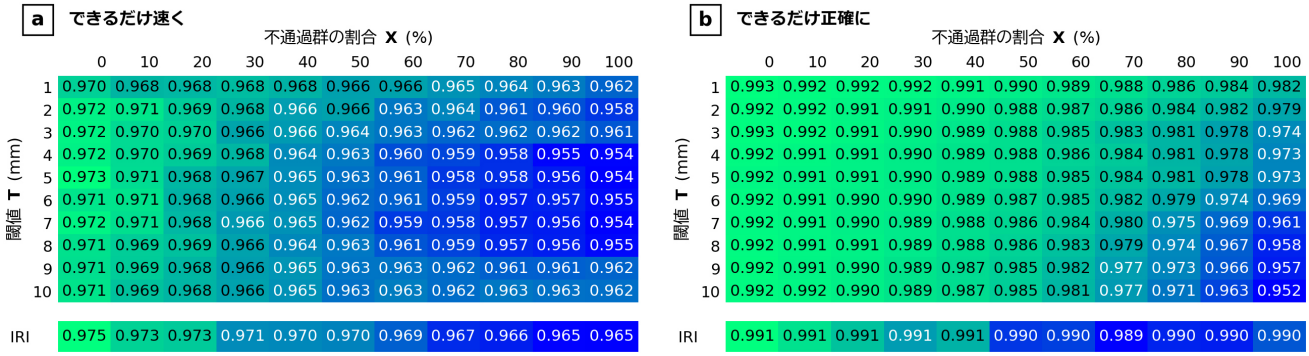

図 9 MT について式 (1) に対するモデル適合度

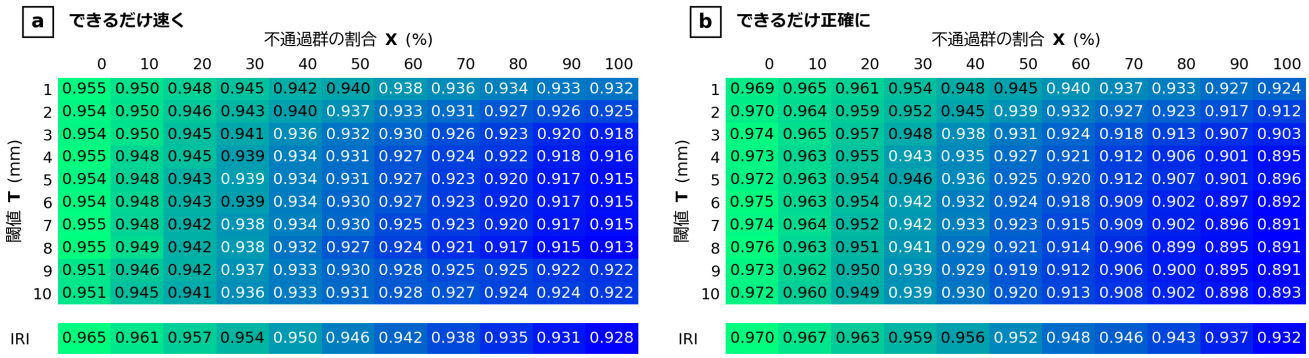

図 10 ER について式 (2) に対するモデル適合度

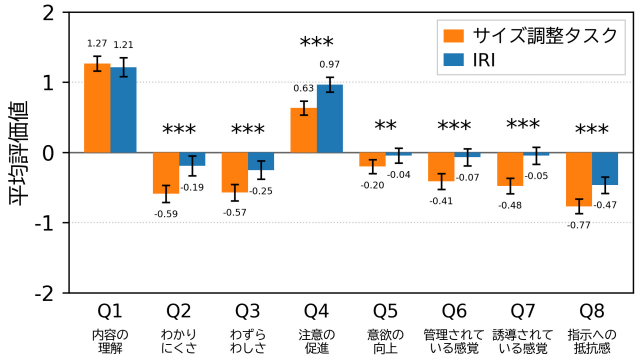

図 11 サイズ調整タスクと IRI の印象評価

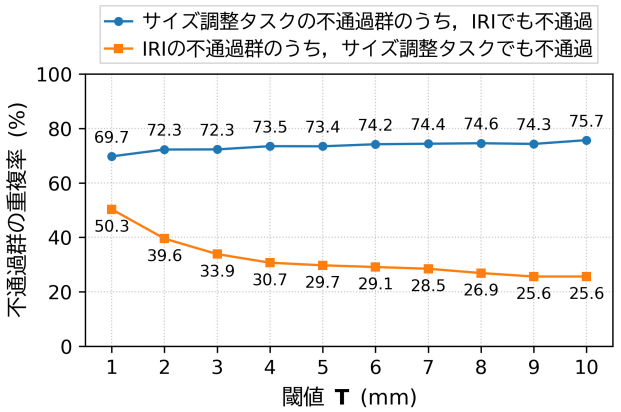

図 12 サイズ調整タスクと IRI の不通過群の重複率

誘導されている感覚（Q7），指示への抵抗感（Q8）といった否定的な印象に関する項目で評価値が有意に低かった．一方で提案手法は，注意の促進（Q4）に加えて意欲の向上（Q5）についても，評価値が IRI より有意に低かった．

以上のことから，IRI と比較しても，提案手法は参加者が感じるわずらわしさや指示への抵抗感に配慮してスクリーニングを行う際の選択肢となると考えられる．一方，本実験では注意の促進だけでなく意欲の向上でも IRI が高く評価されたことから，参加者が感じる負担とタスクへの注意や意欲に対する評価の両面を踏まえて，実験の目的に応じた評価手法を選択する必要があると考えられる．否定的な感情を抱かれにくいが，肯定的にも捉えられにくいという点が提案手法の限界であるといえる．

5.2.3 スクリーニングの対象となる参加者の一致度

サイズ調整タスクと IRI のそれぞれで不通過と判定された参加者の重複率を分析した結果を図 12 に示す．ここでは，サイズ調整タスクの不通過群のうち IRI でも不通過となった割合は，$T = 1, 5, 10$ mm でそれぞれ 69.7%，73.4%，75.7%であった．一方，IRI の不通過群のうちサイズ調整タスクでも不通過となった割合は，同じ閾値でそれぞれ 50.3%，29.7%，25.6%であった．また，Jaccard 係数は同じ閾値でそれぞれ 0.413，0.269，0.237 であり，Cohen の κ は $T = 1$–10 mm の範囲で 0.058–0.075 と低かった．

以上より，IRI においても提案手法とは評価の観点が異なる可能性があるため，評価手法の選択にあたってその手法の評価の観点を考慮する必要性がここでも示唆された．

© 2026 Information Processing Society of Japan

6. まとめ

本稿では，GUI 操作を伴うオンライン実験における参加者スクリーニングのための評価手法として，提案手法であるサイズ調整タスクと既存手法である IMC および IRI を比較した．その結果，いずれの手法でも特に ER モデルにおいてスクリーニングによる適合度の向上が期待できることが示された．また，提案手法は既存手法よりもわずらわしさなどの否定的な印象が少ない一方，注意の促進に関する評価も低く，提案手法の特徴が明らかとなった．さらに，各手法で不通過と判定される参加者の一致度は低く，メインタスクの評価指標であるモデル適合度以外の観点も踏まえて評価手法を選択する必要性が示唆された．

判定の一致度が低かった要因として，各手法で要求される行動の違いによって，評価の観点が異なることが考えられる．サイズ調整タスクでは，GUI を操作してずれを微調整することが求められるため，評価の観点として微調整をする丁寧さの側面が強いと考えられる．一方 IMC や IRI では，文章中の例外的な指示に気づき，それを実行することが求められるため，評価の観点として例外的な指示に気づく注意力の側面が強いと考えられる．こうした評価の観点の違いが，判定の一致度の低さに関係した可能性がある．

また丁寧さや注意力は，メインタスクであるポインティングタスクで求められる行動にも関係する可能性がある．具体的には，丁寧さは特に正確さ重視の条件でターゲット内を正確にタップすることに，注意力は教示の内容を把握してその変化に対応しながら操作を変えることに役立つと考えられる．以上のことから，いずれの手法でもモデル適合度の向上が期待できるものの，メインタスクで重視する行動によって，適した評価手法は異なる可能性がある．そのため，丁寧な操作や教示の変化への対応など，スクリーニングを通じてどのような側面を確認したいのかを明確にし，それに対応する評価の観点を持つ手法を選択することが重要であると考えられる．今後はポインティングタスク以外のメインタスクにも対象を広げ，タスクの特性によって各評価手法を用いたスクリーニングの効果が異なるかも検証する予定である．

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] Miyama, T., Nakamura, S. and Yamanaka, S.: Improving Data Quality via Pre-Task Participant Screening in Crowdsourced GUI Experiments, *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, No. 1009, pp. 1–22 (2026).

[4] 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).

[5] Gummer, T., Roßmann, J. and Silber, H.: Using instructed response items as attention checks in web surveys: Properties and implementation, *Sociological Methods & Research*, Vol. 50, No. 1, pp. 238–264 (2021).

[6] Silber, H., Roßmann, J. and Gummer, T.: The issue of noncompliance in attention check questions: False positives in instructed response items, *Field Methods*, Vol. 34, No. 4, pp. 346–360 (2022).

[7] 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).

[8] 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).

[9] 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).

[10] 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).

[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] 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).

[13] 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).

[14] 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*, No. 317, pp. 1–13 (2023).

[15] 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).

[16] 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).

[17] 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*, No. 579, pp. 1–18 (2023).

© 2026 Information Processing Society of Japan