修士論文

オンラインアンケート回答時の
スマートフォン画面操作状況に基づく不適切回答検出

後上 正樹
奈良先端科学技術大学院大学
先端科学技術研究科
情報理工学プログラム

主指導教員: 安本 慶一
ユビキタスコンピューティングシステム研究室（情報科学領域）

令和 3 年 3 月 17 日提出

本論文は奈良先端科学技術大学院大学先端科学技術研究科に
修士 (工学) 授与の要件として提出した修士論文である。

後上 正樹

審査委員：

主査　　安本 慶一　（情報科学領域　教授）
　　池田 和司　（情報科学領域　教授）
　　荒川 豊　　（九州大学　システム情報科学研究院　教授）
　　松田 裕貴　（情報科学領域　助教）

オンラインアンケート回答時のスマートフォン画面操作状況に基づく不適切回答検出*

後上 正樹

内容梗概

アンケートにおいて，なるべく楽に早くタスクを完了しようとする「Satisficing（努力の最小限化）」という態度により，結果の信頼性が低下する問題がある．より正確な結果を得るためには，Satisficing を検出して分析対象から除外するなどの前処理が必要となる．これまでに，回答時間に基づく検出手法や，指示違反や矛盾を問う質問群を追加する手法が考案されてきた．しかし，前者では回答時間を故意に水増しした回答を適切に除外することができない．また，後者は回答者を疑ってスクリーニングするようなものであり，回答者のモチベーションを損ねて Satisficing を助長してしまう原因となる．これより，スクリーニング質問を用いず，かつ堅牢な Satisficing 検出手法が求められる．先行研究では，回答結果から生成した特徴量を用いて機械学習による検出を試みた結果，55.6%という検出率が報告されているが，実用性の観点からは十分な検出率とは言えない．そこで，回答中の画面操作を利用することで，より高精度に Satisficing が検出可能になるのではないかと考えた．しかしながら，(1) 回答操作が記録可能なアンケートシステムが存在せず，(2) また，記録できたとしてどのような特徴が Satisficing と関連しているのかが不明であった．そこで，世界中で利用されているオンラインアンケートシステム LimeSurvey 用の回答操作記録プラグインを開発し，多人数（5,692 人）の回答操作データを収集して機械学習による不適切回答検出を行なった．Leave One Out 交差検証による評価の結果，85.9%の検出率を達成し，同様のタスクに取り組んだ先行研究の検出率 55.6%を大幅に上回る結果となった．また，

*奈良先端科学技術大学院大学 先端科学技術研究科 修士論文, 令和 3 年 3 月 17 日.

本研究で新たに提案した特徴量の中では，スクロールに関連する特徴量の寄与率が高いことが明らかになった．

キーワード

オンラインアンケート，努力の最小限化，不適切回答検出，機械学習，タッチジェスチャー，スマートフォン

Inappropriate response detection based on smartphone screen operation in online survey*

Masaki Gogami

Abstract

In questionnaires, there is a problem of careless responses due to the attitude of "Satisficing," which is the attempt to complete a task as quickly and easily as possible. In order to obtain results that are closer to the facts, detecting Satisficing and taking some processes such as excluding the responses with Satisficing from the analysis targets are required. Then, variety of methods have been devised to detect Satisficing. One of the method detects it based on response duration and other methods detect it by adding questions that check violations of instructions or inconsistencies. However, the former method cannot properly excludes samples whose response time is deliberately padded. The latter approach is like screening respondents based on suspicion, and these may cause respondents to lose their motivation and even make them Satisficing. Therefore, a robust method for detecting Satisficing without using screening questions is required. In a previous study, although a detection rate of 55.6% was reported as a result of attempting detection by machine learning using features generated from the response results, this is not a sufficient detection rate from the viewpoint of practicality. Therefore, in this thesis, it was thought that a supervised ML model with higher detection rate could be constructed by using answering behavior on screen as features. However, (1) there is no questionnaire system that can record screen operations, and (2) even if it can, it is unclear what features of answering behavior are associated with

*Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 17, .

Satisficing. We developed an answering behavior recording plug-in for LimeSurvey, online questionnaire system used all over the world, and collected a large number of data (5,692 people) in Japan. Then, a variety of features were examined and generated from answering behavior, and we constructed a ML model detecting careless responses. As a result of evaluation by Leave One Out cross-validation, the detection rate of careless response was 85.9%, which is much higher than the detection rate of the previous study. Among the various features we proposed, we found that scrolling (speed and length) in particular contributed to the detection of careless responses.

Keywords:

Online Survey, Satisficing, Inappropriate answer Detection, Machine learning, Touch Gesture

目 次

図 目 次

表 目 次

1. 序論

オンラインアンケートは社会科学分野の調査研究や民間企業のマーケティング活動等に用いられている．クラウドソーシングサービスとの相性が良く，紙ベースのアンケートよりも手軽かつ低コストで大量かつ広範囲に回答を依頼できる利点があり．2015 年からは総務省が実施している国勢調査にも用いられ始めた [1]．しかし，アンケート調査では回答者が必ずしも適切に回答するとは限らない．Maniaci ら [2] は，オンラインアンケートにおける努力の最小限化を後述する様々な手法を用いて調査し，不適切な回答がデータの質や検定力の維持に悪影響を及ぼすことを示している．Simon ら [3] は，人間は何らかの目的を達成するための認知的努力を最低限に抑えるという認知心理学の概念を考案した．Krosnic ら [4] は，これをアンケート調査の領域に適用し，人間が回答要求に対する努力を最小化しようとする傾向を “Satisficing” と定義した．この Satisficing を検出することができれば，不適切な回答に対して適当な処理を施すことで，より真実に近い知見を得ることができると考えられる．

そこで，Oppenheimer ら [5] は IMC（Instructional Manipulation Check），Maniaci ら [2] は ARS（Attentive Responding Scale）および DQS（Directed Question Scale）という Satisficing 検出手法を考案した．これらの手法は検出用の質問をオリジナルの質問票に追加して Satisficing を検出するもので，Satisficing 関連の研究で一般的に使用されている [6, 7, 8]．しかし，追加する必要がある質問は回答者をテストするような内容であるため，挿入することで疑われているように感じるなど，回答者の心理的負荷が増加する可能性がある．これにより，適切に回答している回答者のモチベーションを低下させ，Satisficing を発生させてしまう原因となり得るため，このような質問を挿入することは望ましくない．さらに，Pei ら [9] らは，IMC と DQS の質問に自動で回答するディープラーニングモデルを報告した．これは，上述の Satisficing 検出手法の信頼性を損なったと言える．

尾崎ら [10] は，Satisficing 指標を設置することなく不適切回答を検出するために，機械学習を用いた Satisficing に基づく不適切回答の検出を試みた．対象とした回答用端末は PC であり，回答結果から生成できるデータを特徴量に用いた．様々な機械学習アルゴリズムを試した結果，不適切回答の検出率が最も高かった

ブースティングアルゴリズムの検出率が 55.6%であったと報告されている．いくつものアルゴリズムを試した結果としておおよそ 40%後半 ∼50%前半の検出率となっている点から，精度向上を阻むボトルネックはアルゴリズムではなくデータの質である可能性が考えられる．

また，近年オンラインアンケートの回答に用いられる端末として，スマートフォンが PC に取って代わってきている．これを受けて，Roger ら [11] は，PC/タブレットとスマートフォンでアンケート結果の質にどのような変化があるのかを調査した．この際，評価基準としたのは回答時間，未回答率および連続同一回答数である．結果として，スマートフォンは PC およびタブレットに比べて回答時間が長い傾向が観察された．しかし，結果の信頼性についてはどの端末についても特に差はないと結論づけている．

そこで本研究では，今後さらにオンラインアンケートの回答に利用されることが見込まれるスマートフォンの回答操作に着目し，Satisficing 指標に基づく不適切回答を高精度に検出することを目標とする．しかしながら,（1）回答操作が記録可能なアンケートシステムが存在せず,（2）また，記録できたとしてどのような特徴が Satisficing と関連しているのか不明であった．そこで，我々は世界中で利用されているオンラインアンケートシステム LimeSurvey 用の回答操作記録プラグインを開発し [12]，多人数（5692 人）の様々な回答操作ログを収集して，機械学習による Satisficing 検出を行なった．Leave One Out 交差検証による評価の結果，検出率は 85.9%を達成し，同様のタスクに取り組んだ先行研究の検出率 55.6%を大幅に上回るものとなった．また，本研究で新たに提案した特徴量の中では，スクロールに関連する特徴の寄与率が高い結果となった．

さらなる検出率の向上を目指し，特徴量生成に用いる回答操作データを拡張したモデルを構築してデータの質と量の関係について検証した．結果として，さらなる検出率の向上は実現できなかったが，ページ数（質問数）に従って検出精度も向上することが明らかになった．また，ページ数が少ない場合でも，検出率が大幅に低下することはないため，質問票を作成する際は不要な質問を追加する必要がないことが示唆された．

本論文の構成は次の通りである．2 章で関連研究について紹介する．3 章で Sat-

isficing と関係があると考えられる回答操作について述べる．4 章では，3 章で検討した回答操作を記録するアンケートシステムについて述べる．5 章で Satisficing 指標の検討および不適切回答の定義について説明し，開発したシステムを用いた事前実験により 3 章で考案した回答操作から生成できる特徴量の有用性について述べる．6 章では，その特徴量を用いて機械学習による不適切回答検出の結果および考察を述べる．最後に，7 章で本稿のまとめと今後の展望を述べる．

2. 関連研究

Oppenheimer ら [5] は，アンケート実施前の教示が回答者に伝達されているかどうかを確認することで Satisficing を検出する IMC（Instructional Manipulation Check）という手法を考案した．また，IMC を用いて 2 つの大学の学生（それぞれ 213 人，144 人）を対象として調査を行なった結果，それぞれ 46%，35%が IMC に違反した不適切回答であったと報告されている．三浦ら [13] が日本で実施した調査では，2 社のクラウドソーシングサービス上で各 1800 人を対象としたアンケートにおいて，それぞれ 51.2%，83.8%が IMC に違反した不適切な回答であったと報告されている．このように，適切な認知的コストが払われなかった回答は世界的に一定数存在し，導かれる意思決定を誤ってしまう原因となる可能性が考えられるため問題視されている [14] [15].

この問題に対して，Maniaci ら [2] は，ARS（Attentive Responding Scale）および DQS（Directed Question Scale）という Satisficing 検出手法を考案した．これらの手法は，本来調査したい内容ではないいくつかの質問をアンケートに組み込んで使用する．ARS には Inconsistency と Infrequency という 2 種類がある．Inconsistency は，内容は同じだが文章が微妙に変更された質問対に対する回答の差分に注目するものである．11 の質問対に対する差分の合計が 11 以上であれば Satisficing であると定義されている．Infrequency は，常識的に誰もが選択すると想定される選択肢が存在する質問を設け，その想定選択肢と実際に選択された選択肢の差分に注目するものである．11 問の差分の合計が 12 以上であれば Satisficing であると定義されている．また，DQS はリッカート形式の選択肢の文章中で回答指示（ど

の選択肢を選択させるか，あるいはどの選択肢も選択させない等）を与え，その指示に従わなかった場合 Satisficing であると判断する．三浦ら [16] は，これらの Satisficing 検出手法を効率よく，かつ正確に検出するために，ARS と DQS で挿入しなければならない複数の質問から最低限必要なものを絞り込むことを試みた．しかしながら，当該実験結果においては適切に絞り込むことは難しいと述べられている．

増田ら [17] は，IMC と SC（Seriousness Check: 真面目に回答したかどうかを 2 択で直接的に尋ねる方法）を，不適切回答の検出力の観点から比較した結果，IMC の方が優れていると結論付けた．また，Satisficing を防止するために，冒頭宣誓（回答前に真面目に回答することを宣誓するかどうかを問う設問）を新たに導入した．この設問に対し「真面目に回答する」と答えた群は一般的なリッカート形式質問における連続同一回答数や中間回答数（リッカート尺度の中間に位置する「どちらでもない」のような選択肢を選択する回数）が有意に小さかったと報告されている．このように質問票に工夫を施して Satisficing を検出する手法が検討されているが，これらの手法で挿入する質問はいわばひっかけ問題のようなものであるため，疑われているように感じるなど，回答者の心理的負荷を増加させる可能性がある．そうなると，適切な回答者のモチベーションを低下させ，Satisficing を発生させてしまう原因となり得るため，このような質問を挿入することは望ましくない．さらに，Pei ら [9] は，IMC と DQS を自動で正解するディープラーニングモデルを構築し，75.65%の精度を報告していることから，これらの手法自体の信頼性が脅かされている．

そこで，検出用の質問を必要としない Satisficing 検出手法を模索するために，深井ら [18] は，ページ数，質問数，文字数，スクロール速度，読速度などから，不適切回答であると決定づける回答時間の閾値を算出し，閾値以下の所要時間の回答を除去する手法を提案した．ただし，スクロール速度や読速度など個人に依存する要素は回答者によらず定数としており，事前実験によって計測した被験者群の平均値を用いている．この手法によって閾値以下の回答時間のサンプルを除外した結果，連続同一回答の含有率が 0.66 倍に減少したと報告されている．また，尾崎ら [10] は PC によって回答されたアンケート結果について，機械学習を

用いて Satisficing の検出を試みた．性別，年齢，回答時間，連続同一回答数，ハマラノビス距離，ハマラノビス距離の p 値という，回答結果から得られるデータを特徴量として用いて様々な機械学習モデルを適用し，最も検出率が高いモデルで 55.6%という結果を報告した．本研究では，特徴量としてスマートフォンの回答操作に関するデータを用いることで，不適切回答の検出率を向上させることを目指す．なお，本研究で回答に用いる端末をスマートフォンに限定した理由は，オンラインアンケートの回答に用いられる端末としてスマートフォンが PC に取って代わってきている背景があるためである [19]．Roger ら [11] は，PC，タブレットおよびスマートフォンにおいて，アンケートの回答の質にどのような変化があるのかを調査した．この際，評価基準としたのは回答時間や未回答率および連続同一回答数である．結果として，スマートフォンは PC およびタブレットに比べて回答時間が長い傾向が観察された．一方で，結果の信頼性についてはどの端末についても特に差はないと結論づけており，スマートフォンによる回答増加が回答の質についてネガティブでないことが示されている．

3. Satisficingに関連するデータ

不適切回答検出に寄与するデータを検討し，本稿で扱う項目を表 1 に示す.「質問形式」欄では，各データがリッカート形式もしくは自由記述形式のどちらの質問形式に対応するのかを表している．なお，同欄の項目「全体」は質問形式に関係なくアンケート全体に関する回答操作データであることを表す．また,「独自に追加」欄では，不適切回答検出のために用いるデータとして，以前から考案されていた – 印で示すデータに加えて本研究独自で考案し追加したものを⋆印で表す.

表 1 Satisficing 検出に用いる回答操作データ

回答操作データ	単位	質問形式	独自に追加
リッカートの回答時間	s	リッカート	–
自由記述の回答時間	s	自由記述	–
選択肢の変更回数	回	リッカート	⋆
テキストの削除回数	回	自由記述	⋆
スクロール長	px	全体	⋆
スクロール時間	s	全体	⋆
スクロール速度	px/s	全体	⋆
逆スクロール回数	回	全体	⋆
非操作時間が長すぎる回数	回	全体	⋆
連続同一回答数	問	リッカート	–
中間回答数	問	リッカート	–
文字数	文字	自由記述	–

回答時間は，これまでの研究でも不適切回答検出のために用いられてきた．アンケート調査会社等でも,「アンケート全体の回答時間」が短すぎるサンプルを調査結果から除外する例がある [20]．このようなフィルタリングに引っかからない不適切回答者や，アンケートのある部分のみ不適切な回答をする回答者なども存在し得るが，依然として回答時間は Satisficing に強く関係すると考える．本稿で

は「リッカートの回答時間」と「自由記述の回答時間」に分け，それぞれの形式の質問に対する平均回答時間とした．

「選択肢の変更回数」,「テキストの削除回数」,「テキストの削除率」は，正確に回答しようと考えている状態が表れる回答操作データであると考えられる．そのため，少なくとも雑な回答をしようとした場合には発生しないと推測する．したがって，これらの回答操作データも Satisficing に関連すると考える．テキストの削除率は，削除回数を文字数で除して算出される．テキストの削除は文字数に伴って発生確率が変わるため，純粋な削除回数よりも不適切回答に寄与することが期待される．

「スクロール長」は，一回のスクロール操作による画面移動量と定義した.「スクロール時間」も同様に一回のスクロール操作にかかった時間とし,「スクロール速度」は「スクロール長」を「スクロール時間」で除算した値とした．これらは質問間の移動という行動を表す一つのパラメータとして捉えることができる．Satisficing が発現している場合は，早く終わらせたい思いから質問間の移動が粗くかつ速くなると考えられる．

「逆スクロール回数」は，100px 以上の逆向きのスクロールを 1 回と定義した．100px とした理由は，LimeSurvey のアンケートを一般的なスマートフォンで回答する際に前の質問に戻るために最低限必要な移動量であるためである．アンケートの回答中に前の質問の回答を変更する場合や，ページ冒頭の質問文を読み直す際の行動を表していると考えられる．このような行動は丁寧に回答しようとしている状態を裏付けるものであるため，逆スクロール回数がほとんどないような場合は Satisficing が発現している可能性があると考えられる．

「非操作時間が長すぎる回数」は，画面に触れていない時間が基準値以上の回数であり，5.3 節で説明する事前実験 1 の結果を基に本回答操作データを考案した．この基準値はリッカート形式では 10 秒，自由記述形式では 40 秒と定義した．この基準値を上回る非操作時間は，回答にかかるであろう想定時間の範囲を超えているため，何か他の作業をしながら回答している状態であると見なす．ながら操作は質問への注意を逸らす要因であるため，本回答操作データの値が大きい場合は Satisficing が発現している可能性が高いと考えられる．

「最大連続同一回答数」は，リッカート形式の質問において同じ選択肢を連続で回答する回数の最大値である．Satisficing でない場合でも同一回答になることはあるが，Satisficing が発現している回答者はその最大値が大きくなるとされている [21].

「中間回答数」は，リッカート形式の質問において中間の「どちらでもない」のような選択肢を選択する回数である．これに関しても Satisficing でない状態でも中間回答を選択する場合はある．一方で，Satisficing が発現しており，自分の意見を確認して表明するという認知的コストを払わず実質的に回答を放棄するような場合に，中間の選択肢を選択する傾向がある [22]．そのため，中間回答数は Satisficing と関連すると考える．

「文字数」は，自由記述形式の質問 1 問あたりの文字数とする．質問文で文字数や内容の具体度の指定がない場合，回答者は一文で回答する場合と，数文に渡って具体的に回答する場合がある．この差が Satisficing に関連しており，文字数少ない方が Satisficing 傾向が強いであろうと推測される．

また，下記リストに示すいくつかの回答操作データに対する変動係数，回答者内および回答者間の偏差も有効な特徴量になり得ると考えた．変動係数とは，回答者ごとに，標準偏差を平均値で除した値である．各特徴量のばらつきを特徴量とする際に，回答者ごとに異なる平均値の個人差を吸収するために標準偏差ではなく変動係数を用いた．回答を進めるにつれて不注意な回答が増加するという傾向 [23] を考慮し，アンケート冒頭との差が不適切回答の検出に寄与するのではないかと考えた．回答者間の偏差は，ある回答者の値と回答者全体の平均との差である．これは，回答者全体の平均的な回答行動との違いを表しているため，不適切回答の検出に寄与するのではないかと考えた．

- 変動係数：スクロール長，スクロール時間，スクロール速度，リッカートの回答時間
- 回答者間の偏差：選択肢の変更回数，テキストの削除回数，テキスト文字数，スクロール長，スクロール速度，逆スクロール回数
- 回答者内の偏差 : テキストの削除回数，テキストの削除率，スクロール速度

4. 回答操作記録アンケートシステム

3 章で述べた特徴量を記録するためのアンケートシステムを開発した [12]．4.1 節では，システム構成の検討の流れを説明し，開発したシステムの構成について述べる．4.2 節では，本システム内の回答操作記録部分であるプラグインが記録するデータについて述べる．

4.1 システム構成の検討

3 章において議論した特徴量を生成するための回答操作ログを記録するシステムの構成として，下記のような方法が考えられた．

- OS 側でのタッチジェスチャー取得
- カメラによる動画撮影
- アンケートアプリケーションの自作
- ヒートマップ可視化ツール
- 既存アンケートシステムの活用

これらそれぞれの利点と欠点について検討し，本稿で提案するシステムの採用理由を述べる．

OS 側でタッチジェスチャーを取得する例として，Touch Analyzer [24] がある．これは Android 端末の OS が出力するイベントデバイスファイルを逆解析することで，使用中のアプリケーションに依存せずタッチジェスチャーを取得する．この手法であれば特別なアプリケーションのインストールが不要であるが，スマートフォンを PC と USB で接続する必要があり，広範なアンケートの実施においてスケーラビリティの問題が発生する．また，Android 限定であり，iOS では使用できないため汎用性に欠ける点も本研究において不適であると考えた．

カメラで回答操作を録画して，ムービーの解析によって回答操作を抽出するアプローチも考えられる．この手法もアプリケーションに非依存であるが，一方でアンケートシステムとは別にカメラを用意する環境が必要となり，Touch Analyzer 同様，スケーラビリティの観点で問題がある．

専用のアンケートアプリケーションを全て自作するという方法も考えられた．自作するため設計および実装の自由度が高く，理論上どの OS についても所望の挙動を実現することができる．また，回答者がアプリケーションをインストールするという以外の特殊なセッティングは不要であり，スケーラビリティの問題もクリアできる．一方で，アンケート実施者は広く使われている既存のアンケートシステムではなく，操作に慣れないアンケートアプリケーションを導入し，回答者に案内する必要がある．この点は，実施者側におけるデメリットとなり，スケーラビリティが良くない．また，筆者についてもアンケートシステムをゼロから自作するのは骨が折れる作業であり，本来フォーカスすべき目的ではなく手段に時間と労力を割くことになりかねないため，この方法は得策とは言い難い．

Web ページの改善のためにページ上でのユーザの行動をヒートマップ等で可視化する ClickTale [25] などのサービスについても検討した．しかしながら，汎用性が高い反面，アンケートに特化した回答操作データの記録はできない点がデメリットとして挙げられた．

既存の Web アンケートシステムとして広く使われているものに，Google Form，SurveyMonkey，LimeSurvey がある．これらを用いて機能要件を実現できれば，ゼロからアプリケーションを作る必要はない．このうち Google Form と SurveyMonkey はサービスとして運用されており，利用者が自作のプログラムを組み込む余地がなく，要件を満たすシステムが実現できない．一方で，LimeSurvey はオープンソースであり，プラグインとして独自に開発した機能を組み込む仕組みがある．本研究で必要となる JavaScript でタッチジェスチャーや回答内容を取得する技術を導入することもできる．

これら各アプローチの利点と欠点を考慮し，要件の実現性，システムの拡張性や拡散性の観点から，本研究では LimeSurvey をベースとしたシステムを提案することとした．LimeSurvey は，Google Form [26] や Survey Monkey [27] 等とは異なり，JavaScript を用いて独自のプラグインを作成することが可能である．我々はこの仕組みを用いて，表 1 に示す回答操作データを取得するプラグイン（以降，Operation Logger と呼ぶ）を独自に開発し，回答操作が記録可能なアンケートシステムを構築した．本システムの概観を図 1 に示す．Operation Logger の導入方

法は，LimeSurvey をホスティングするアプリケーションサーバ上に JavaScript と PHP のファイルを配置し，サービス内の質問設定画面で JavaScript ファイルの読み込み設定をするのみである．これにより，標準機能で記録される回答結果データと共に回答操作データがデータベースに格納される．なお，回答者側はソフトウェアの追加や設定の変更が一切必要ないため，高い有用性を有するシステムとなっている．

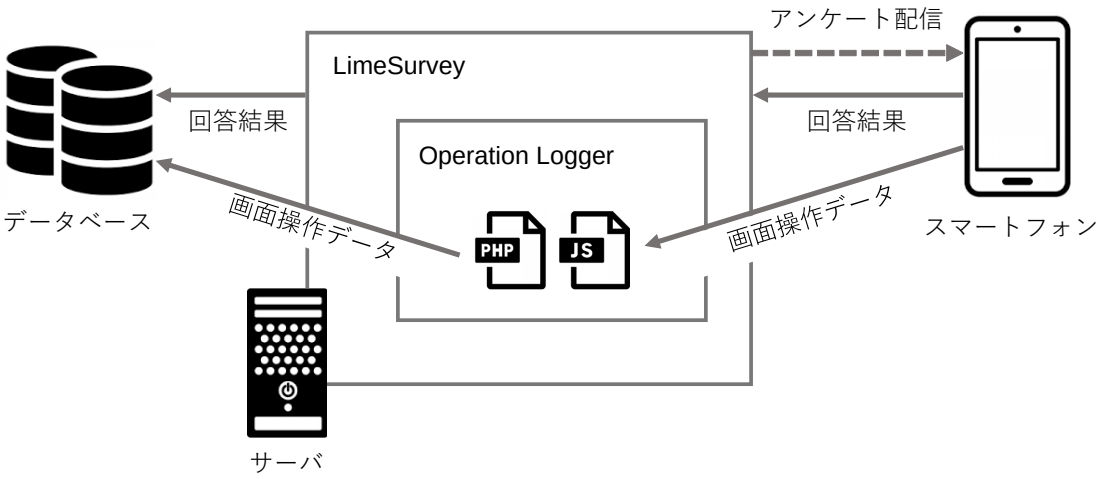


図 1 回答操作記録アンケートシステムの概観

4.2 Operation Logger で記録するデータ

Operation Logger においてデータを取得するイベントは大別するとタッチイベント，選択肢のタップ，テキストの入力の 3 種類である．

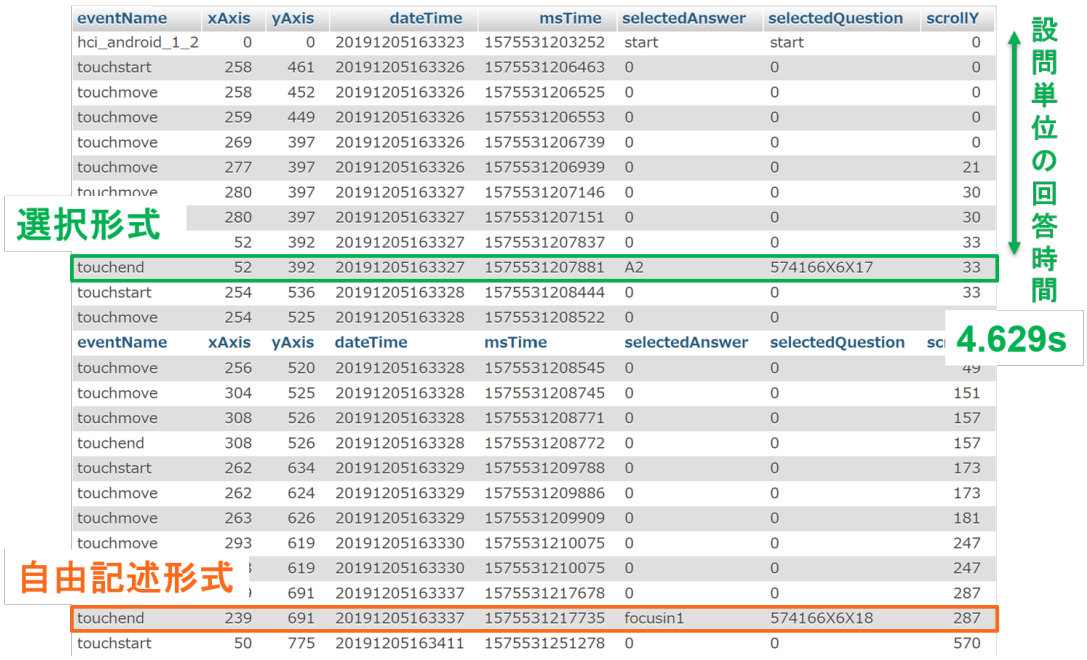
タッチイベントの種類は，touchstart，touchmove，touchend の 3 種類であり，それぞれスクリーンが指を検知した瞬間，指がスクリーン上で動いている間，指がスクリーンを離れた瞬間のタイミングで発火する. これらのタイミングにおける時刻，画面内の座標，ページ上端からの移動量，タッチイベントの種類を取得し，データベースに格納する．これにより,「スクロール長」,「スクロール時間」,「ス

クロール速度」,「逆スクロール回数」,「非操作時間が長すぎる回数」を検出することができる．

選択肢をタップしたタイミングでは，回答時刻，質問の ID，選択肢の ID を取得する．これにより，リッカート形式の「質問単位の回答時間」と「選択肢の変更」を検出することができる．

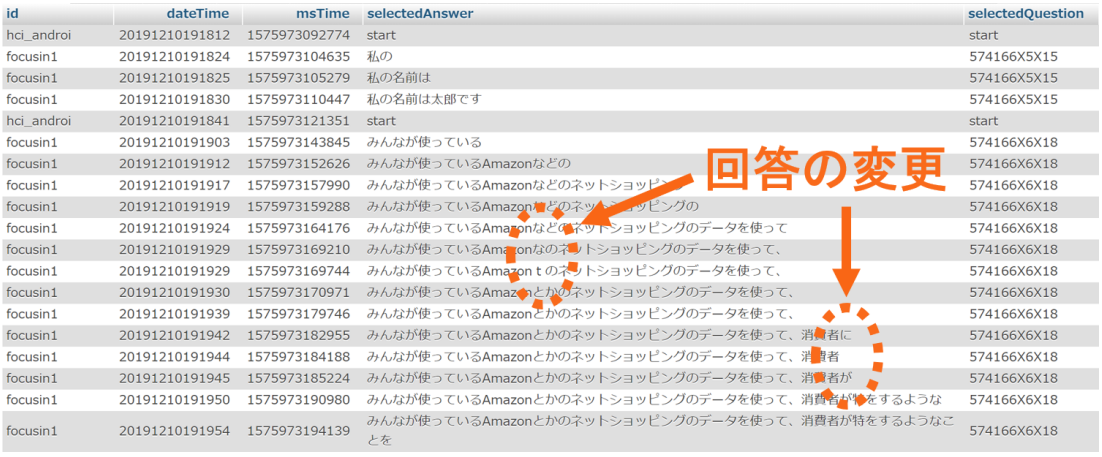
テキストを入力したタイミングでは，回答時刻，質問の ID，入力内容，レコード生成トリガの種類を取得し，データベースに格納する．これにより，自由記述形式の「自由記述の回答時間」,「テキストの削除回数」,「テキストの削除率」が検出できる．テキスト入力中の記録単位を検討した結果，1 レコードを生成するタイミングは「入力のない時間が 1 秒経過」,「デリートから入力への切り替わり」,「入力からデリートへの切り替わり」,「フォーカスアウト」とした．

これらのデータが記録されているデータベースのメインテーブルとテキストテーブルのスクリーンショットを図 2，3 に示す．図 2 中では，リッカート形式，自由記述形式それぞれの質問に回答した際に挿入されるレコードの例を示す．また，ページの開始時点から 1 問目の回答までの時間を基に，設問単位の回答時間を取得する例を示す．図 3 中では，自由記述形式の回答中にテキストの変更をした場合に記録されるデータの例を示す．



eventName	xAxis	yAxis	dateTime	msTime	selectedAnswer	selectedQuestion	scrollY
hci_android_1_2	0	0	20191205163323	1575531203252	start	start	0
touchstart	258	461	20191205163326	1575531206463	0	0	0
touchmove	258	452	20191205163326	1575531206525	0	0	0
touchmove	259	449	20191205163326	1575531206553	0	0	0
touchmove	269	397	20191205163326	1575531206739	0	0	0
touchmove	277	397	20191205163326	1575531206939	0	0	21
touchmove	280	397	20191205163327	1575531207146	0	0	30
touchmove	280	397	20191205163327	1575531207151	0	0	30
touchmove	52	392	20191205163327	1575531207837	0	0	33
touchend	52	392	20191205163327	1575531207881	A2	574166X6X17	33
touchstart	254	536	20191205163328	1575531208444	0	0	33
touchmove	254	525	20191205163328	1575531208522	0	0	
eventName	xAxis	yAxis	dateTime	msTime	selectedAnswer	selectedQuestion	sc
touchmove	256	520	20191205163328	1575531208545	0	0	49
touchmove	304	525	20191205163328	1575531208745	0	0	151
touchmove	308	526	20191205163328	1575531208771	0	0	157
touchend	308	526	20191205163328	1575531208772	0	0	157
touchstart	262	634	20191205163329	1575531209788	0	0	173
touchmove	262	624	20191205163329	1575531209886	0	0	173
touchmove	263	626	20191205163329	1575531209909	0	0	181
touchmove	293	619	20191205163330	1575531210075	0	0	247
touchmove	[illegible]	619	20191205163330	1575531210075	0	0	247
touchmove	[illegible]	691	20191205163337	1575531217678	0	0	287
touchend	239	691	20191205163337	1575531217735	focusin1	574166X6X18	287
touchstart	50	775	20191205163411	1575531251278	0	0	570

図 2 取得するデータの例（メインテーブル）



id	dateTime	msTime	selectedAnswer	selectedQuestion
hci_androi	20191210191812	1575973092774	start	start
focusin1	20191210191824	1575973104635	私の	574166X5X15
focusin1	20191210191825	1575973105279	私の名前は	574166X5X15
focusin1	20191210191830	1575973110447	私の名前は太郎です	574166X5X15
hci_androi	20191210191841	1575973121351	start	start
focusin1	20191210191903	1575973143845	みんなが使っている	574166X6X18
focusin1	20191210191912	1575973152626	みんなが使っているAmazonなどの	574166X6X18
focusin1	20191210191917	1575973157990	みんなが使っているAmazonなどのネットショッピ	574166X6X18
focusin1	20191210191919	1575973159288	みんなが使っているAmazonなどのネットショッピングの	574166X6X18
focusin1	20191210191924	1575973164176	みんなが使っているAmazonなどのネットショッピングのデータを使って	574166X6X18
focusin1	20191210191929	1575973169210	みんなが使っているAmazonなどのネットショッピングのデータを使って、	574166X6X18
focusin1	20191210191929	1575973169744	みんなが使っているAmazonとかのネットショッピングのデータを使って、	574166X6X18
focusin1	20191210191930	1575973170971	みんなが使っているAmazonとかのネットショッピングのデータを使って、	574166X6X18
focusin1	20191210191939	1575973179746	みんなが使っているAmazonとかのネットショッピングのデータを使って、	574166X6X18
focusin1	20191210191942	1575973182955	みんなが使っているAmazonとかのネットショッピングのデータを使って、消費者に	574166X6X18
focusin1	20191210191944	1575973184188	みんなが使っているAmazonとかのネットショッピングのデータを使って、消費者	574166X6X18
focusin1	20191210191945	1575973185224	みんなが使っているAmazonとかのネットショッピングのデータを使って、消費者が	574166X6X18
focusin1	20191210191950	1575973190980	みんなが使っているAmazonとかのネットショッピングのデータを使って、消費者が特をするような	574166X6X18
focusin1	20191210191954	1575973194139	みんなが使っているAmazonとかのネットショッピングのデータを使って、消費者が特をするようなことを	574166X6X18

図 3 取得するデータの例（テキストテーブル）

5. 事前実験

4 章で述べた回答操作記録プラグイン Operation Logger を用いて，提案する特徴量と Satisficing との関係を調査するための事前実験を 2 つ実施した．5.1 節でアンケート内容について述べる．5.2 節では，Satisficing 指標および，それに基づく本研究における不適切回答の定義について述べる．5.3 節では，考案した特徴量と Satisficing との関係を調査するために研究室内の学生を対象に実施したアンケート（事前実験 1）について述べる．5.4 節では，普段アンケートが実施されている環境であるクラウドソーシング上でアンケートを実施し（事前実験 2），考案した特徴量について不適切回答群と適切回答群の間に統計的な差があるのかを検証した．なお，本研究におけるアンケートは奈良先端科学技術大学院大学人を対象とする研究に関する倫理審査委員会の承認を受けて実施した（承認番号 : 2020-I-2）.

5.1 アンケート内容

アンケート実験で用いた質問票の概略を図 4 に示す．この質問票は三浦ら [16] が公開している質問票 [28] のうち，Big5 尺度，自尊感情尺度，認知欲求，アンケートへのモチベーションおよび DQS，ARS から成るリッカート形式部分をベースとし，次の 3 点を変更した．1 点目は，後述する 1~6 および 15~17 ページの追加である．2 点目は，ARS の質問対を回答者に悟られにくくするためにダミー質問を 11 問追加した点である．3 点目は，DQS の質問が 5 ページ連続してページ末尾に配置されていたため，DQS の質問箇所を悟られないために 3 問に減らし，ページ末尾や冒頭を避けて配置した点である．最終的に，全 17 ページ，128 問（5 段階リッカート形式 124 問，自由記述形式 4 問）で，回答目安時間が約 15 分の質問票とした．

各ページについて説明する．1~3 ページは回答者の基本属性を取得するための質問である．4 ページ目は回答者 ID をプラグインのシステムと共有するためのものである．5 ページ目は回答者ごとのスクロール操作のベースラインを計測するための質問である．例として，ベースライン計測ページとメイン質問ページの実際のアンケート画面のスクリーンショットを図 5 に示す．回答者は下方向に

スクロールして回答を進める．質問内容は，図 5（a）のように，ある選択肢を選択するように指示するものである．一般的な質問よりも認知的コストが低く，Satisficing が発現しにくい状態でのスクロール行動を計測する．これは，後述する特徴量である回答者内の相対的スクロール速度を算出するために計測した．6 ページ目は回答者ごとに，自由記述形式質問における 1 文字あたりの削除回数で表される削除率のベースラインを計測するための質問である．指定した文章を入力する際の削除回数を文字数で除算した値を，各回答者の削除率のベースラインとした．これは 3 章で述べたテキストの削除回数および削除率の回答者内偏差（delete_num_Selfdev，delete_rate_Selfdev）を算出するために計測した．7 ～14 ページ目は心理状態を評価する質問票等を用いた．そのうち，8, 10, 12 ページ目に DQS の質問を，9, 11 ページ目に ARS の質問を配置した．DQS と ARS については次節で説明する．15～17 ページ目では，簡単な自由記述形式の質問を設けた．なお，メインの内容である 7 ページ目以降は必須回答設定を OFF とした．また，自由記述形式の文字数も一つの特徴量であるため，文字数指定もなしとした．

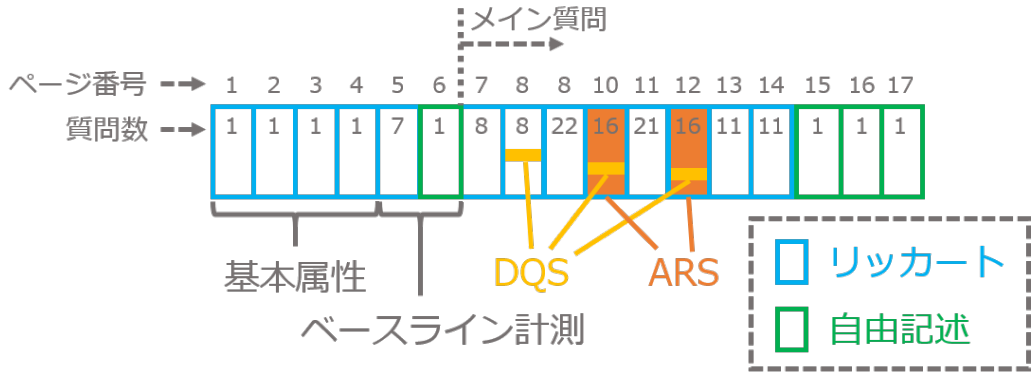


図 4 質問項目の概略

(a) ベースライン計測ページ　　(b) メイン質問ページ

図 5 実際のアンケート画面のスクリーンショット

5.2 不適切回答

本研究では,「不適切回答」を DQS（Directed Question Scale）と ARS（Attentive Responding Scale）の 2 つの Satisficing 指標に基づいて定義した．指標を 2 つ用いたのは，これらの指標が異なる視点から Satisficing を検出するためである．

DQS [2] は，回答の指示文を質問と同列に設置し，その指示に従わなかった場合 Satisficing であるとする指標である．本研究では，3 問の DQS を設置し，1 問以上指示に反した回答をした場合に Satisficing であると定義した．ARS [2] には，Inconsistency と Infrequency という 2 種類がある．Inconsystency は，内容が同じで文章を微妙に変更した質問対に対する回答の差分に注目するものである．11 の質問

対に対する差分の合計が 11 以上であれば Satisficing であるとされる．Infrequency は，常識的に誰もが選択すると想定される選択肢が存在する質問を設け，その想定選択肢と実際に選択された選択肢の差分に注目するものである．11 問の差分の合計が 12 以上であれば Satisficing であるとされる．本研究では，Inconsistency と Infrequency のどちらか 1 つでも Satisficing であるサンプルを ARS 全体についての「Satisficing」とした．本研究ではこれらの指標を基に，正解ラベルとなる「不適切回答者」を「DQS，ARS 両方の指標で Satisficing である回答者」と定義した．

なお，この他にしばしば用いられる IMC（Instructional Manipulation Check）[5] は，アンケートの回答前に受ける特別な教示を回答者が正確に認識しているかどうかをチェックするものである．本研究で用いた質問票には特に特別な教示を要する質問を含んでいないため，IMC を Satisficing 指標として利用できないと判断し，使用しなかった．

5.3 事前実験 1

Operation Logger の簡単な動作確認と，記録される回答操作データから生成できる特徴量と Satisficing の関係性を検討するために本実験を実施した．

5.3.1 実験手続き

．研究室内の学生 7 名に上述のアンケートの回答を依頼し，各自のスマートフォンを用いて回答してもらいデータを収集した．動作確認の正解データを記録するために，回答時の画面周辺の映像を録画した．

回答結果を解析したところ，DQS については Satisficing 回答が 0 件であり，ARS についても Inconsistency の違反が 2 件あったのみであった．この結果からは Satisficing と回答操作の関係を分析することができないため，Satisficing が発現するように追加で同様のアンケート実験を 2 回実施することとした．2 回目の実験のねらいは，1 回目と同じアンケートに回答しなくてはならないという面倒さによって，Satisficing が発現しやすい回答環境を構築することであった．また，3 回目の実験では，回答完了後に 500 円相当のクオカードを付与する旨を伝え，外発的な

動機付けによって Satisficing が発現しにくい回答環境の構築を図った．まとめると，1 回目の実験は 2 回目の面倒さを与えるためのダミー実験とし，Satisficing が発現しやすいと想定した 2 回目と，発現しにくいと想定した 3 回目を主な分析対象とした．

5.3.2 結果

まず，Operation Logger の動作確認については，記録した動画と回答操作データを照合し，問題なく記録出来ていることを確認した．

各実験の回答結果を基に，DQS・ARS のクラスごとの該当者数を算出した結果を表 2 に示す．DQS に関しては，DQS 3 問のうち全問指示に従った場合は「not_Satisficing_DQS」，1 問以上指示に従わなかった場合は「Satisficing_DQS」というラベルを割り当てた．ARS は Inconsistency と Infrequency の 2 種類あるため，どちらも Satisficing でないサンプルを「not_Satisficing_ARS」，どちらか 1 つが Satisficing であるサンプルを「Satisficing_ARS_OR」，どちらも Satisficing であるサンプルを「Satisficing_ARS_AND」というラベルに割り当てた．

閾値に基づいた Satisficing の数は 3 回の実験を通して大きな変化はなかった．また，DQS については Satisficing の回答者が存在しなかった．ARS については，3 つの実験の結果を図 6 に示す．横軸が Inconsistency，縦軸が Infrequency の値を表しており，赤い点線はそれぞれの Satisficing と判定する閾値である．また，サンプルの右肩の数字は何回目の実験かを表し，色が同じサンプルは同じ回答者の結果を表す．Satisficing の数としては大きな違いはないものの，回答者単位で実験 2 と 3 の結果を比較すると，2 回目の方が 3 回目よりも値が大きい傾向が見られた．

表 2 Satisficing 指標の各クラスの回答数（研究室内事前実験）

Satisficing 指標	クラス	クラスの説明	回答数 [件]		
			1 回目	2 回目	3 回目
DQS	not_Satisficing_DQS	DQS が Satisficing でない群	7	7	7
	Satisficing_DQS	DQS が Satisficing である群	0	0	0
ARS	not_Satisficing_ARS	ARS が Satisficing でない群	5	4	5
	Satisficing_ARS_OR	ARS のどちらか一つが Satisficing である群	2	3	2
	Satisficing_ARS_AND	ARS の両方が Satisficing である群	0	0	0

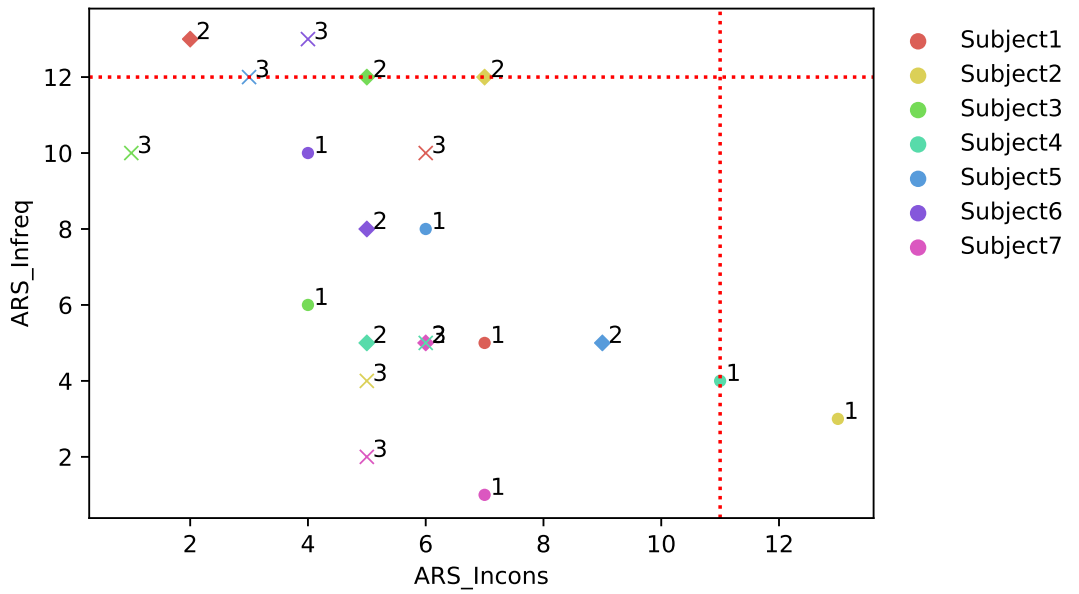


図 6 事前実験 1 の ARS の結果

そこで次に，実験 2 と 3 の差の観点から，ARS と各特徴量の関係をプロットした．特に ARS と相関があるように見られた特徴量について図 7，8 に示す．横軸が 2 種類の ARS について，実験 2 から実験 3 を引いた値である．つまり，実験 2 の方が ARS が大きくなるという想定に沿った差は正の値となる．縦軸は順に，スクロール速度の回答者間偏差，リッカートの回答時間について同じく実験 2 から

実験 3 を引いた値を表す．これらの図では，点線で区切られたエリアのうち，第一象限に属するサンプルが多い．つまり，ARS の増加に伴って各特徴量が増加していることを表すため，これらの特徴量は不適切回答の検出に寄与する可能性が高いと考えられる．

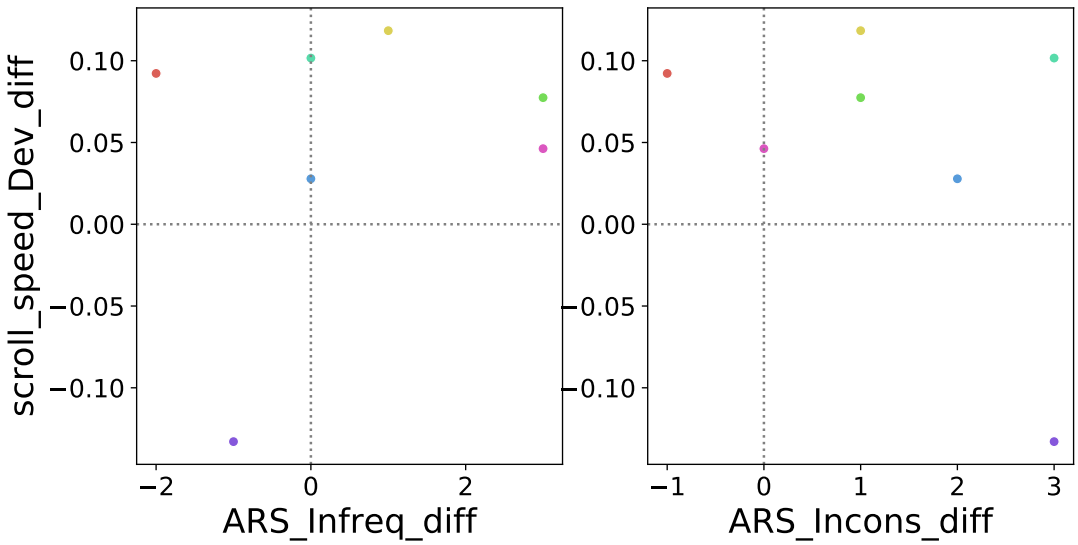


図 7 スクロール速度の回答者間偏差と ARS の実験 2，3 間の差

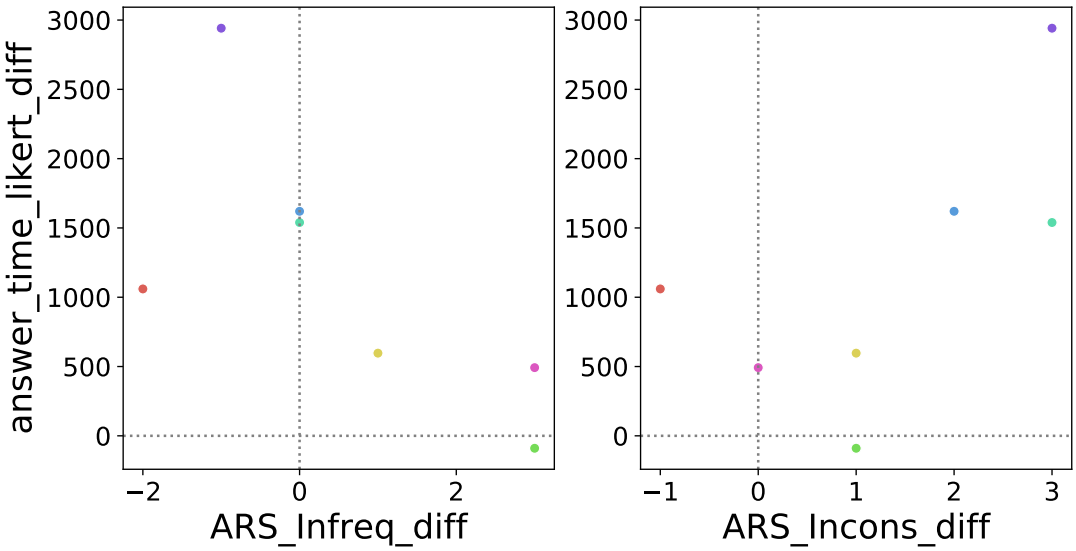


図 8 リッカートの回答時間と ARS の実験 2，3 間の差

次に，全回答者の質問ごとの回答時間を図 9 に示す．横軸が質問 ID を表しており，左から順に実際のアンケートの表示順となっている．縦軸は各質問の回答に要した時間を，凡例は回答者 ID を表す．アンケート開始直後に見られるスパイクの原因は，アンケートのスタイルに慣れていないことであると考えられる．その他にも，アンケート中盤で長い時間をかけているケースがある．極端に回答時間が長い質問は，回答内容の検討に時間がかかっているのではなく，アンケート以外の事をする，いわゆるながら操作を行なっていると考えられる．これを基に，3 章で述べた「非操作時間が長すぎる回数」という特徴量が不適切回答の検出に寄与するのではないかと考えた．

図 9 各実験の質問単位の回答時間

5.4 事前実験 2

事前実験 1 では，研究室内の学生を対象としたため，Satisficing である回答が少なく，不適切回答の検出に寄与する特徴量の分析が十分にできなかった．そのため，クラウドソーシングを用いて一般人を対象としたアンケートのデータで，考案した特徴量について不適切回答群と適切回答群の間に統計的な差があるかどうかを検証した．

5.4.1 実験手続き

本実験は，オンラインアンケートが実際に実施されているクラウドソーシングの環境として，Yahoo!クラウドソーシングを用いた．回答者には報酬として，1 回答当たり 5 円相当のポイントを付与した．また，回答に用いる端末はスマートフォンに限定した．本稿で提案する手法はタッチスクリーンが搭載されている端末であれば，タブレットやノート PC にも対応可能であると考えられる．ただし，画面サイズが一定以上大きくなると，質問文や選択肢の画面上の配置が変わるため，それをある程度統一するためにスマートフォンに限定した．また，スマートフォンの中でも画面サイズが多様な機種が存在するが，実用性の観点からも，どのようなスマートフォンにも適用できる Satisficing 検出モデルであることが望ましいため，それ以上の制限は設けなかった．スマートフォンに限定する方法は，クラウドソーシングサイト上のタスク説明欄とアンケートのスタート画面でスマートフォンで回答するように指示する文章を記載した．また，Operation Logger 側で Javascript によって回答に用いた端末を記録し，スマートフォン以外の回答は分析対象から除外した．

まず，事前分析によって収集するサンプルサイズを決定した．検出力は 0.8 とし，正例と負例の比率は，表 5.4.1 に示す三浦ら [16] が報告した値を参考に設定した．これは，本研究において三浦ら [16] が公開している質問票を基に作成したアンケートを用いて実験を行なったためである．本研究の不適切回答者の定義（DQS 違反かつ，ARS Inconsistency または ARS Infrequency 違反）に沿った不適切回答率は計算できないが，本定義で必須となる違反項目 DQS の平均値が約 4%

なので，0.04 とした．

表 3 三浦らが報告した各 Satisficing 指標の違反率

Satisficing 指標	違反率
ARS（Inconsistency）	4.7%
ARS（Infrequency）	9.0%
DQS	2.2∼5.6%

これらの定数を用いて計算したところ，両側検定において必要なサンプルサイズは 818 であった．そこで，無効回答等の数も考慮して事前実験では 1000 人の回答を募集した．

5.4.2 統計検定

本実験では，回答者 1000 人のうち，回答操作データの利用に同意した 817 人のデータを分析対象とした．回答結果を基に，DQS・ARS および正解ラベルについてクラスごとの該当者数を算出した結果を表 4 に示す．正解ラベルは 5.2 節で定義した通り，表 4 の「Satisficing ラベル」欄が「Satisficing_DQS」 かつ，「Satisficing_ARS_OR」または「Satisficing_ARS_AND」であるサンプルを不適切（正例），それ以外のサンプルを適切（負例）としてラベリングした．結果として,「不適切回答（正例）」が 54 件,「適切回答（負例）」が 763 件であった．

表 4 Satisficing 指標および正解ラベルの各クラスの該当者数と割合（クラウドソーシング上事前実験）

Satisficing 指標	クラス	クラスの説明	回答数 [件]	割合 [%]
DQS	not_Satisficing_DQS	DQS が Satisficing でない群	732	90
	Satisficing_DQS	DQS が Satisficing である群	85	10
ARS	not_Satisficing_ARS	ARS が Satisficing でない群	672	82
	Satisficing_ARS_OR	ARS のどちらか一つが Satisficing である群	128	16
	Satisficing_ARS_AND	ARS の両方が Satisficing である群	17	2
正解ラベル	適切	DQS が Satisficing かつ ARS が一つ以上 Satisficing である群	763	93
	不適切	「不適切」以外の群	54	7

これらのデータを基に，各特徴量についてウェルチの t 検定を実施し，不適切回答群と適切回答群の母平均の差を検定した．有意水準は 5%とし，検定は両側検定とした．全特徴量の検定結果を p 値を図 10 に示し，表 5 にまとめる.「大小関係」欄では，適切回答群よりも不適切回答群の平均値が大きいか小さいかを示す.「有意差」欄では，統計的な有意差があるか否かを示している．×印は有意差がないこと，∗ 印は有意差があることを表している．

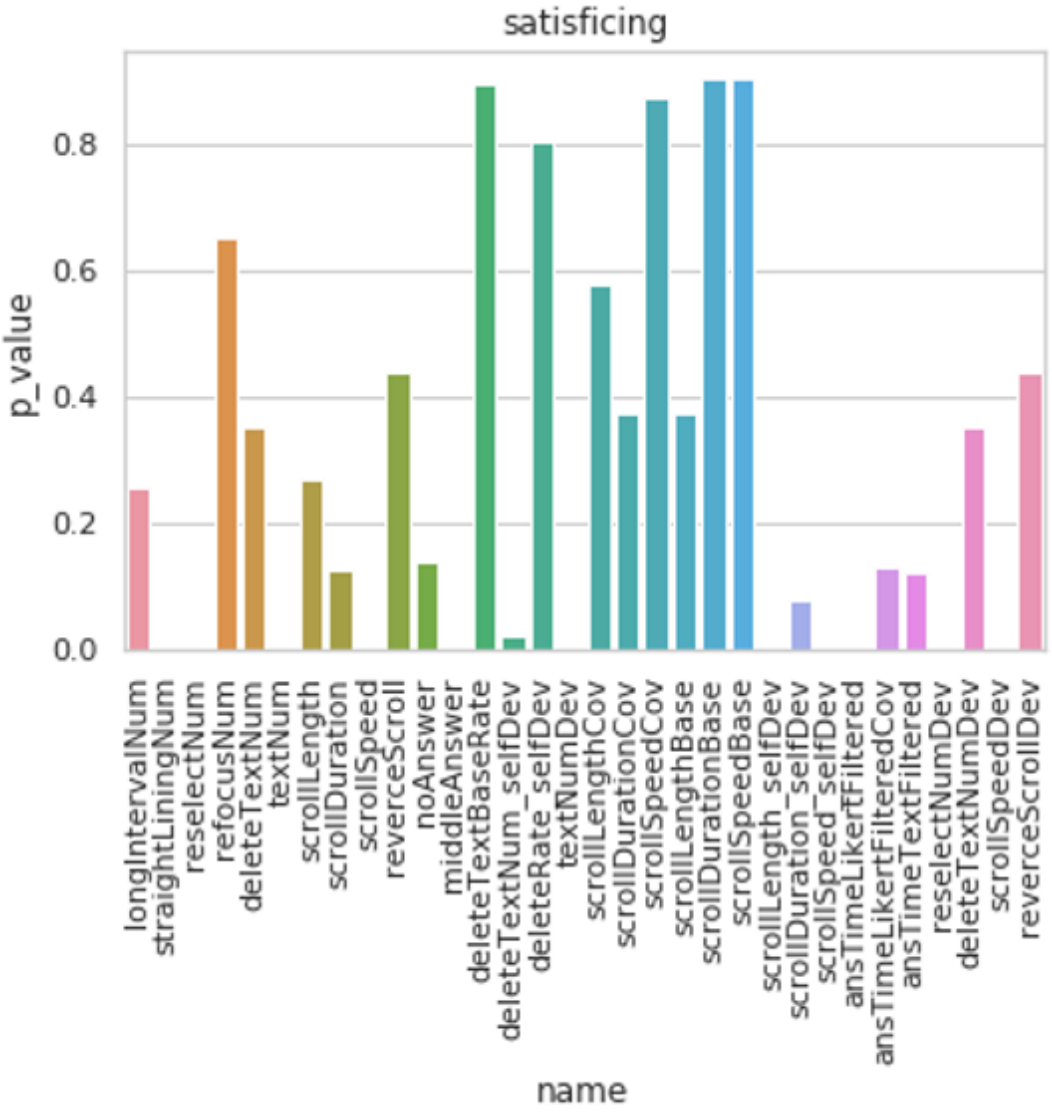


図 10 ウェルチの t 検定の各特徴量の p 値

表 5 適切回答群と不適切回答群の差の検定結果と大小関係

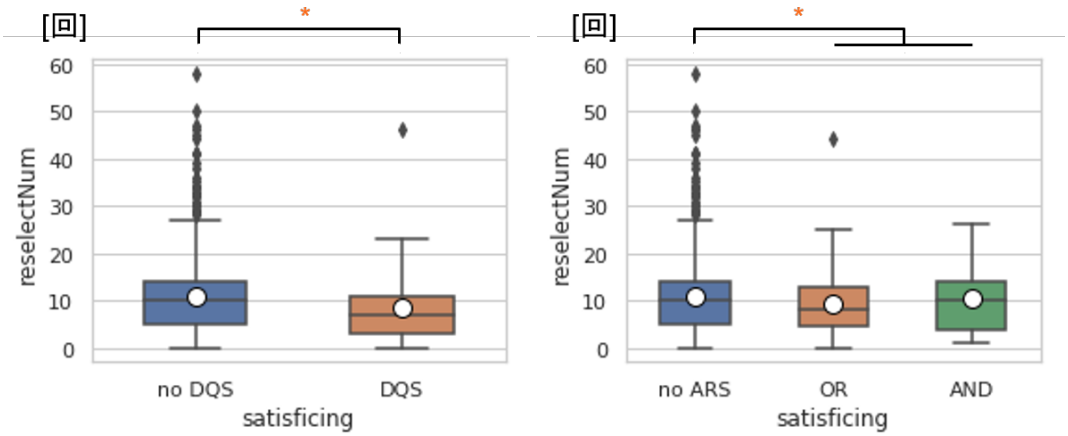
特徴量	大小関係[a)]	有意差[b)]	
		DQS	ARS
リッカートの回答時間	小	＊	＊
自由記述の回答時間	小	−	−
選択肢の変更回数	小	＊	＊
テキストの削除回数	小	−	−
スクロール長	小	−	−
スクロール速度	大	＊	＊
逆スクロール回数	小	−	−
非操作時間が長すぎる回数	小	−	−
最大連続同一回答数	大	＊	＊
中間回答数	大	＊	＊
文字数	小	＊	＊
リッカートの回答時間の変動係数	大	−	＊
スクロール長の変動係数	小	−	−
スクロール速度の変動係数	大	−	−
文字数の偏差	大	−	−
テキストの削除回数の偏差	大	＊	−
選択肢の変更回数の偏差	大	＊	＊
逆スクロール回数の偏差	大	−	−
テキストの削除回数の自己 BL との差	大	＊	−
スクロール長の自己 BL との差	大	＊	＊
スクロール速度の自己 BL との差	大	＊	＊

[a)] 適切回答群に対して，不適切回答群の平均が大きいか，小さいかを示す．
[b)] − 印は有意差がないこと，＊ は有意差があることを表す．

また，有意差が認められた一部の特徴量について，クラスごとの箱ひげ図を図 11 ～図 14 に示す．これらの図において注目すべき部分は，ひげからはみ出てい

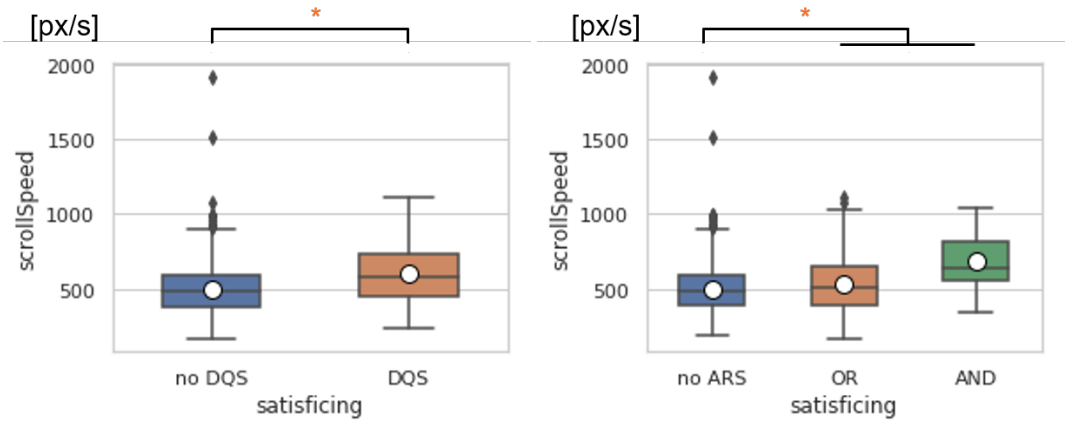
るサンプルであると考える．例として図 14 (a) を見ると，「no DQS」群のひげよりも上部にプロットされたサンプルの値は，「DQS」群にはほとんど存在しない．これら大きく外れた特徴は，次の課題である機械学習モデルによる Satisficing 検出において，重要な特徴となり得るだろう．

各種変動係数は統計的な差がほとんど見られない結果となったことから，各種特徴量のばらつきの大きさは Satisficing 検出において寄与度が低い可能性が示唆された．また，各種回答者内偏差に関しては，おおかた有意な差が認められた．特に，スクロール長およびスクロール速度に関して解釈すると，不適切回答群は一定のスクロール操作で回答を行なっていないと考えられる．



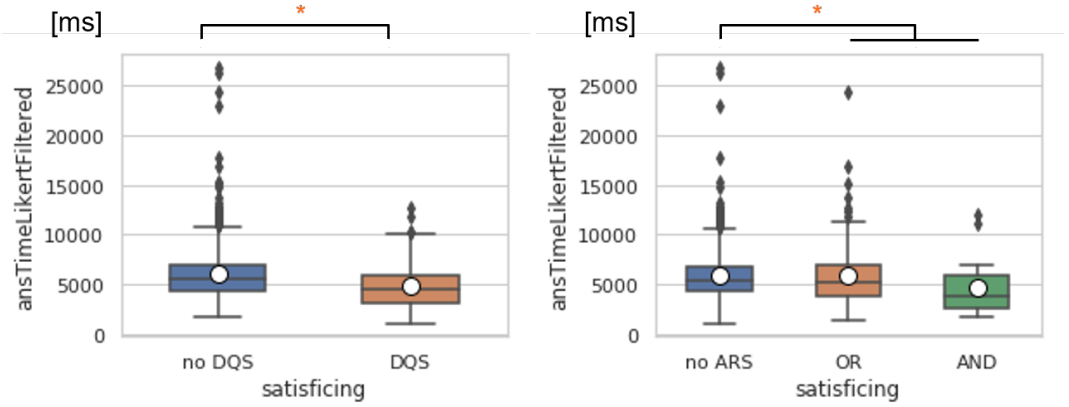
(a) DQS　　(b) ARS

図 11　特徴量「選択肢の変更回数」のクラスごとの箱ひげ図



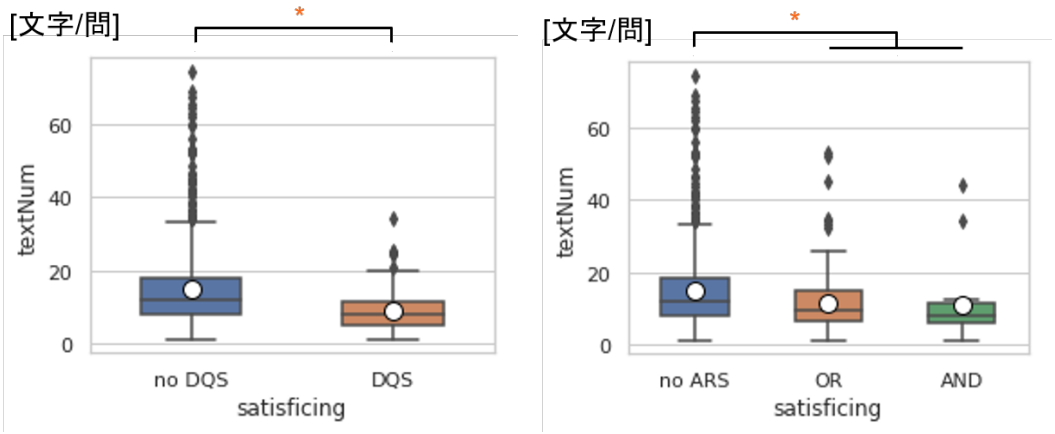
(a) DQS (b) ARS

図 12 特徴量「スクロール速度」のクラスごとの箱ひげ図



(a) DQS (b) ARS

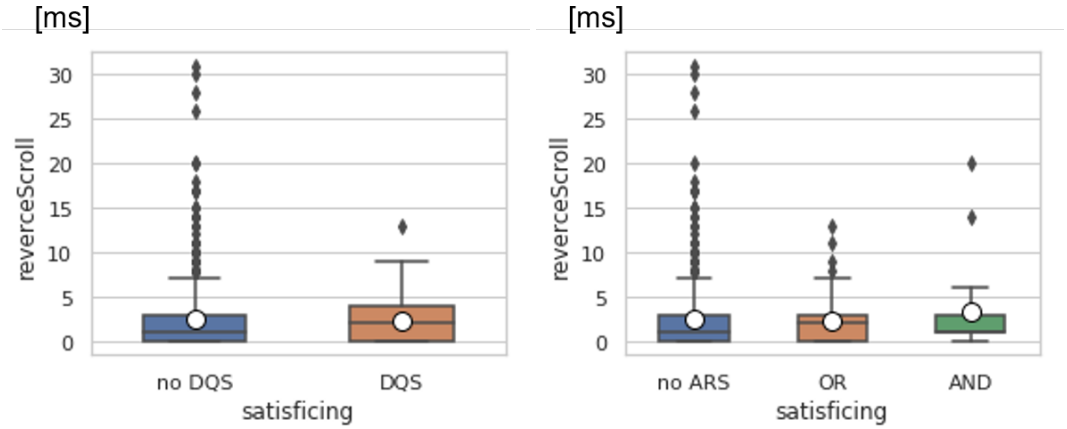
図 13 特徴量「回答時間」のクラスごとの箱ひげ図



(a) DQS　　(b) ARS

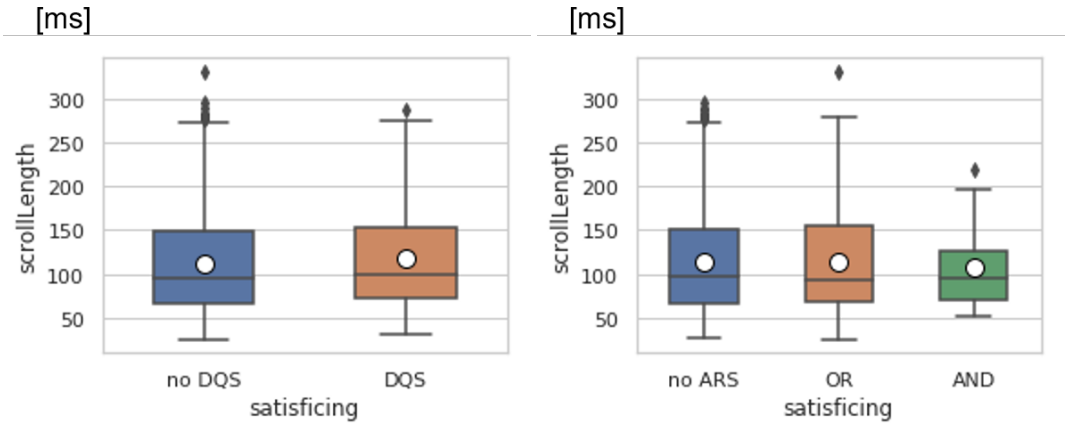
図 14 特徴量「文字数」のクラスごとの箱ひげ図

有意差が見られなかった特徴量の例として，逆スクロール回数とスクロール長の結果をそれぞれ図 15，16 に示す．逆スクロール回数に関しては，Satisficing 状態では少なくなるだろうという仮説に沿った統計的な差を示す結果にはならなかった．しかし，図 15 では DQS，ARS 共にひげよりも大きい値をとるサンプルはやはり適切回答群に属している．これより，統計的には差があったとは言えないものの，先述のように Satisficing を検出する機械学習モデルにおいては重要な特徴量である可能性がある．スクロール長に関しては，絶対値では有意差がなかったにも関わらず，図 17 に示す回答者内偏差では DQS，ARS 共に有意な差が認められた．テキストの削除回数に関しても，表 5 を参照すると回答者内偏差の方が有意な差が見られた（DQS のみ）．これより，絶対的な特徴量よりも回答者個人のベースラインを考慮した特徴量が Satisficing 検出に寄与する可能性が示唆された．



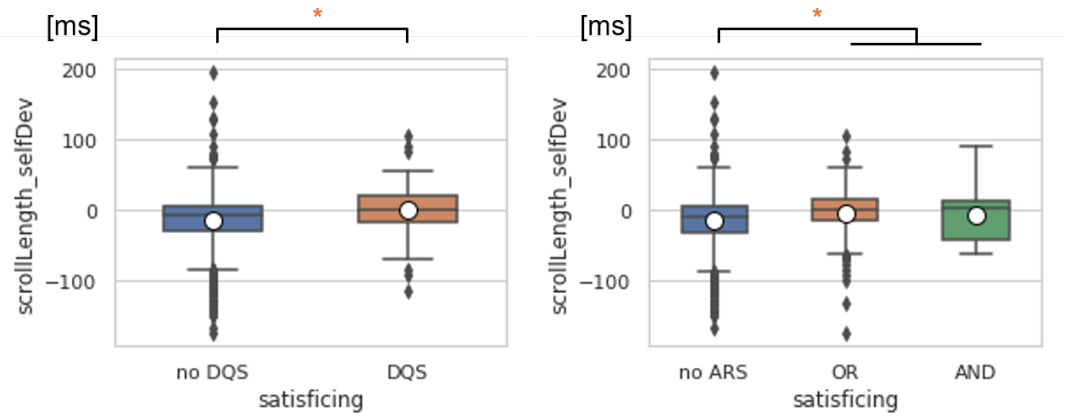
(a) DQS (b) ARS

図 15 特徴量「逆スクロール回数」のクラスごとの箱ひげ図



(a) DQS (b) ARS

図 16 特徴量「スクロール長」のクラスごとの箱ひげ図



(a) DQS　　(b) ARS

図 17 特徴量「スクロール長の回答者内偏差」のクラスごとの箱ひげ図

6. 機械学習による不適切回答検出

5 章で述べた事前実験において，不適切回答群と適切回答群との間に有意な差が認められた特徴量が存在した．これを受けて，同一のアンケートを追加実施してデータを増やし，機械学習による不適切回答の検出を行なった．本研究では，「不適切回答の検出」という課題を「適切 or 不適切の 2 値分類」問題とした．6.1 節では，分類モデルのアルゴリズムおよび検証方法に関して述べる．6.2 節では，特徴量の追加・選択の方法について述べる．6.3 節では，6.2 節の各過程における分類精度について述べる．6.4 節では，分類精度向上を目指して 6.3 節の結果から考えた仮説を検証した結果について述べる．

6.1 基礎データ

本研究では，回答者 5692 人のうち，回答操作データの利用に同意した 4940 人のデータを分析対象とした．前節と同様に，各クラスに対応するサンプル数を表 6 に示す．最終的に正解ラベルは,「不適切（正例）」が 247 件,「適切（負例）」が 4693 件であった．この正解ラベルを機械学習モデルの分類対象とした．

表 6 Satisficing 指標および正解ラベルの各クラスの回答数と割合（本実験）

Satisficing 指標	クラス	回答数 [件]	割合 [%]
DQS	not_Satisficing_DQS	4520	91
	Satisficing_DQS	420	9
ARS	not_Satisficing_ARS	4066	82
	Satisficing_ARS	874	18
正解ラベル	適切	4693	95
	不適切	247	5

また，各特徴量の平均値と標準偏差を表 6.1 に示す.

表 7 各特徴量の平均値と標準偏差

特徴量名	説明	単位	平均値	標準偏差
answer_time_likert	リッカートの回答時間の平均	s	2.68×10^{6}	1.88×10^{8}
answer_time_likert_Cov	リッカートの回答時間の変動係数	–	2.61	1.17
answer_time_text	自由記述の回答時間の平均	s	31738.10	39305.45
reselect_num	選択肢の変更回数の平均	回	10.56	7.77
reselect_num_Dev	reselect_num の回答者間偏差	回	-2.11×10^{-16}	7.77
delete_num	文字の削除回数の平均	回	7.71	12.32
delete_num_Dev	delete_num の回答者間偏差	回	1.54×10^{-14}	12.30
delete_num_Selfdev	delete_num の回答者内偏差	回	-41.27	97.10
delete_rate_Selfdev	文字数に対する文字の削除回数の回答者内偏差	回/文字	-1.24	2.63
scroll_length	スクロール長の平均	px	118.84	62.60
scroll_length_Cov	スクロール長の変動係数	–	0.431	0.129
scroll_length_Selfdev	scroll_length の回答者内偏差	px	-12.67	42.28
scroll_duration	スクロール時間の平均	s	332.18	209.98
scroll_duration_Cov	スクロール時間の変動係数	–	0.891	0.490
scroll_duration_Selfdev	scroll_duration の回答者内偏差	s	-41.70	169.27
scroll_speed	スクロール速度の平均	px/s	513.10	168.51
scroll_speed_Cov	スクロール速度の変動係数	–	0.471	0.155
scroll_speed_Dev	scroll_speed の回答者間偏差	px/s	2.98×10^{-14}	168.33
scroll_speed_Selfdev	scroll_speed の回答者内偏差	px/s	27.15	117.30
reverce_scroll_num	逆向きスクロール回数の平均	回	2.01	2.98
reverce_scroll_num_Dev	reverce_scroll_num の回答者間偏差	回	2.26×10^{-15}	2.98
long_interval_num	非操作時間が長すぎる回数	回	4.86	4.47
straight_lining_num	連続同一回答数の最大値	問	4.77	5.44
middle_answer_num	中間回答数	問	30.53	16.34
text_num	文字数の平均	文字	12.29	10.15
text_num_Dev	text_num の回答者間偏差	文字	-1.15×10^{-14}	10.10

6.2 機械学習モデル

尾崎ら [10] は数ある機械学習モデルの中で，ブースティングアルゴリズムが最も不適切回答の検出率が高かったと報告している．この知見に倣って本研究でも，決定木をベースとした勾配ブースティングアルゴリズムである LightGBM [29] を用いた．ハイパーパラメータのチューニングには，ベイズ最適化アルゴリズムを用いた自動最適化ツールである Optuna [30] を用いた．モデルの精度評価には Accuracy，Precision，Recall，F1 Score を用いた．ただし，本研究の目的は不適切回答の検出であることから，その検出率を表す Recall に注目すべきであると考える．モデルの汎化性能の検証は，対象サンプルを 1 つだけテスト用として交差検証する Leave One Out 交差検証によって行なった．本研究で扱うデータは正例と負例の比率が不均衡であったため，図 18 に示すように負例をランダムにダウンサンプリングし，正例 : 負例 $= 1 : 1$ として評価した．このとき，汎化性能評価の観点から，ダウンサンプリングは 5 回行ない，精度はその平均値とした．

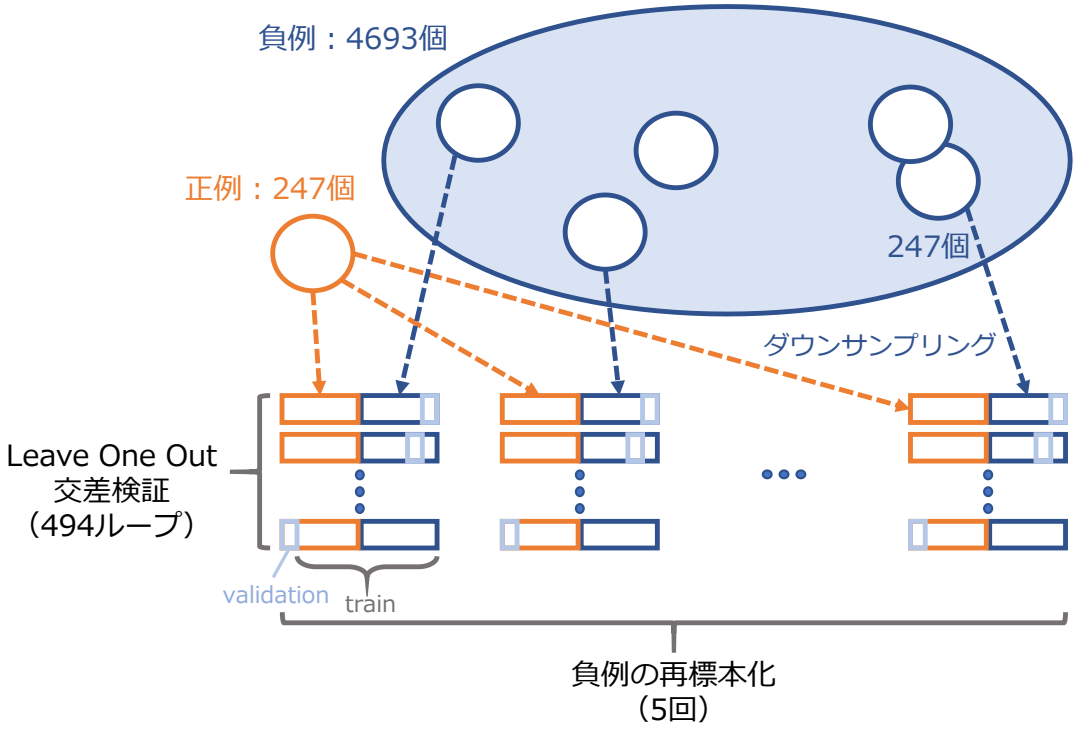


図 18 ランダムダウンサンプリングと Leave One Out 交差検証のデータの分割方法

6.3 特徴量の追加・選択

使用した特徴量の一覧を表 8 に示す．また，特徴量の追加・選択を 3 段階行なった際の各モデルにおける特徴量の使用不使用についても示す．まず，モデル 1 では，表 1 に示した特徴量に加え，変更回数を文字数で除して算出する「テキストの削除率」を追加した．これは，テキストの削除は文字数に応じて発生確率が変わるため，純粋な変更回数よりも寄与率が高くなる可能性が考えられたためである．また，スクロール長，スクロール時間，スクロール速度，リッカートの回答時間の変動係数を追加した．変動係数とは，回答者ごとに，標準偏差を平均値で除した値である．ばらつきを特徴量とする際に，回答者ごとに異なる平均値の個人差を吸収した特徴量とするために標準偏差ではなく変動係数を用いた．

- 変動係数：スクロール長，スクロール時間，スクロール速度，リッカートの回答時間

次に，モデル 1 の特徴量を基に，各種特徴量に対する回答者内および回答者間の偏差を特徴量に追加した（これを「モデル 2」と呼ぶ）．絶回答者間の偏差は，ある回答者の値と回答者全体の平均との差である．回答者内の偏差は，質問票 5，6 ページ目において計測した回答者ごとのベースラインとアンケート全体の平均値との差である．それぞれどの特徴量について追加したのかを下記に示す．

- 回答者間の偏差：選択肢の変更回数，テキストの削除回数，テキスト文字数，スクロール長，スクロール速度，逆スクロール回数
- 回答者内の偏差 : テキストの削除回数，テキストの削除率，スクロール速度

さらに，図 19 に示す特徴量の相関行列で高い相関があるペアについて，片方を削除する特徴選択を行なった（これを「モデル 3」と呼ぶ）．このとき，図 20 に示すモデル 2 における特徴量の寄与率が低い方の特徴量を削除した．該当する特徴量を図中に丸印でマークしている．ここで，削除された特徴量は相対的なものよりも絶対的なものが多いことが確認できる．これより，絶対的な値よりも回答者内・回答者間の相対的な値の方が分類に寄与していると言える．

表 8 特徴量の説明と各モデルでの使用不使用

特徴量名	説明	単位	大小関係[1]	モデル 1	モデル 2	モデル 3
answer_time_likert	リッカートの回答時間の平均	s	小	○	○	○
answer_time_likert_Cov	リッカートの回答時間の変動係数	–	大	○	○	○
answer_time_text	自由記述の回答時間の平均	s	小	○	○	×
reselect_num	選択肢の変更回数の平均	回	小	○	○	×
reselect_num_Dev	reselect_num の回答者間偏差	回	小	×	○	○
delete_num	文字の削除回数の平均	回	小	○	○	×
delete_num_Dev	delete_num の回答者間偏差	回	小	×	○	×
delete_num_Selfdev	delete_num の回答者内偏差	回	大	×	○	×
delete_rate_Selfdev	文字数に対する削除回数の回答者内偏差	回	大	×	○	○
scroll_length	スクロール長の平均	px	大	○	○	○
scroll_length_Cov	スクロール長の変動係数	–	小	○	○	○
scroll_length_Selfdev	scroll_length の回答者内偏差	px	大	×	○	○
scroll_duration	スクロール時間の平均	s	小	○	○	×
scroll_duration_Cov	スクロール時間の変動係数	–	大	○	○	○
scroll_duration_Selfdev	scroll_duration の回答者内偏差	s	小	×	○	×
scroll_speed	スクロール速度の平均	px/s	大	○	○	×
scroll_speed_Cov	スクロール速度の変動係数	–	小	○	○	○
scroll_speed_Dev	scroll_speed の回答者間偏差	px/s	大	×	○	○
scroll_speed_Selfdev	scroll_speed の回答者内偏差	px/s	大	×	○	○
reverce_scroll_num	逆向きスクロール回数の平均	回	大	○	○	○
reverce_scroll_num_Dev	reverce_scroll_num の回答者間偏差	回	大	×	○	×
long_interval_num	非操作時間が長すぎる回数	回	小	○	○	○
straight_lining_num	連続同一回答数の最大値	問	大	○	○	○
middle_answer_num	中間回答数	問	大	○	○	○
text_num	文字数の平均	文字	小	○	○	×
text_num_Dev	text_num の回答者間偏差	文字	小	×	○	○

[1] 各特徴量の平均値について，不適切回答群が適切回答群よりも大きいか小さいかを表す．

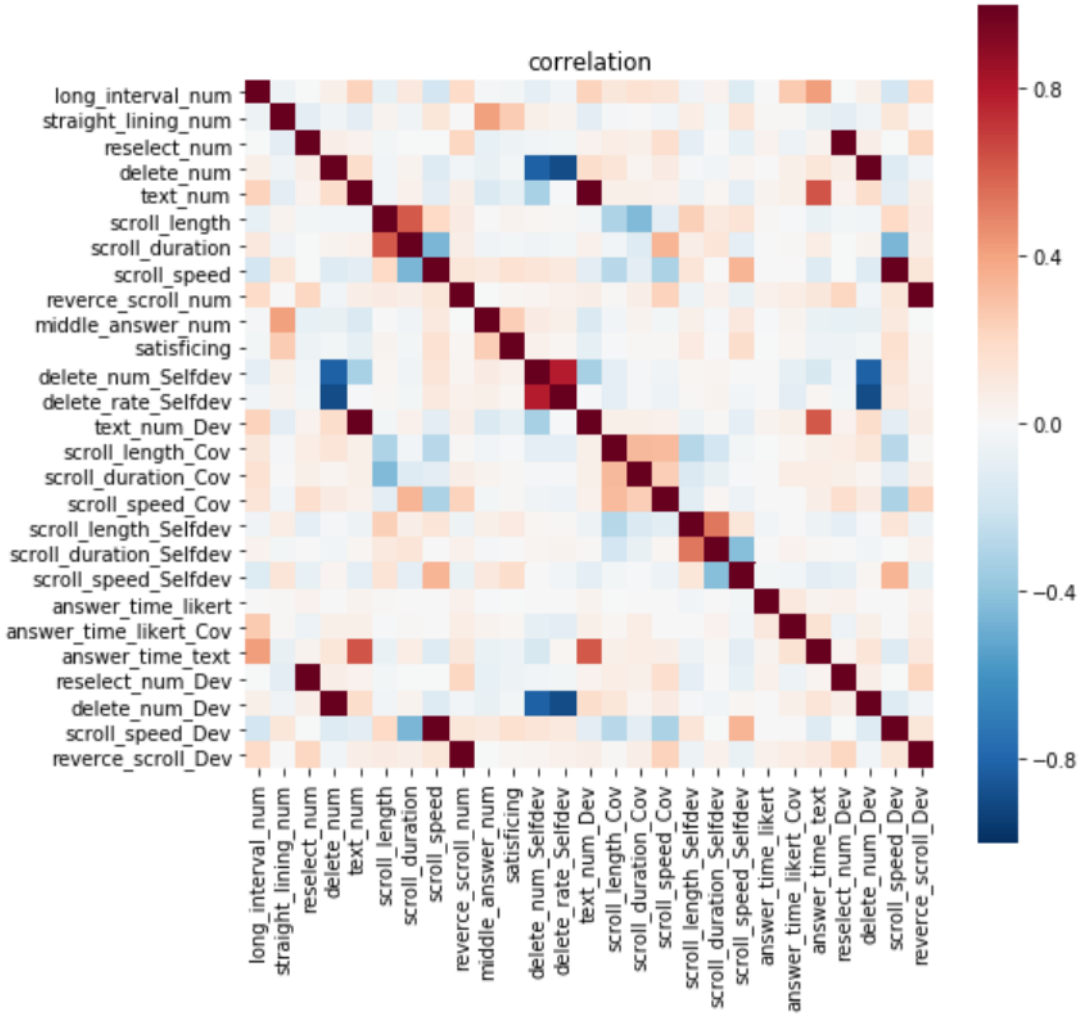


図 19 特徴量同士の相関

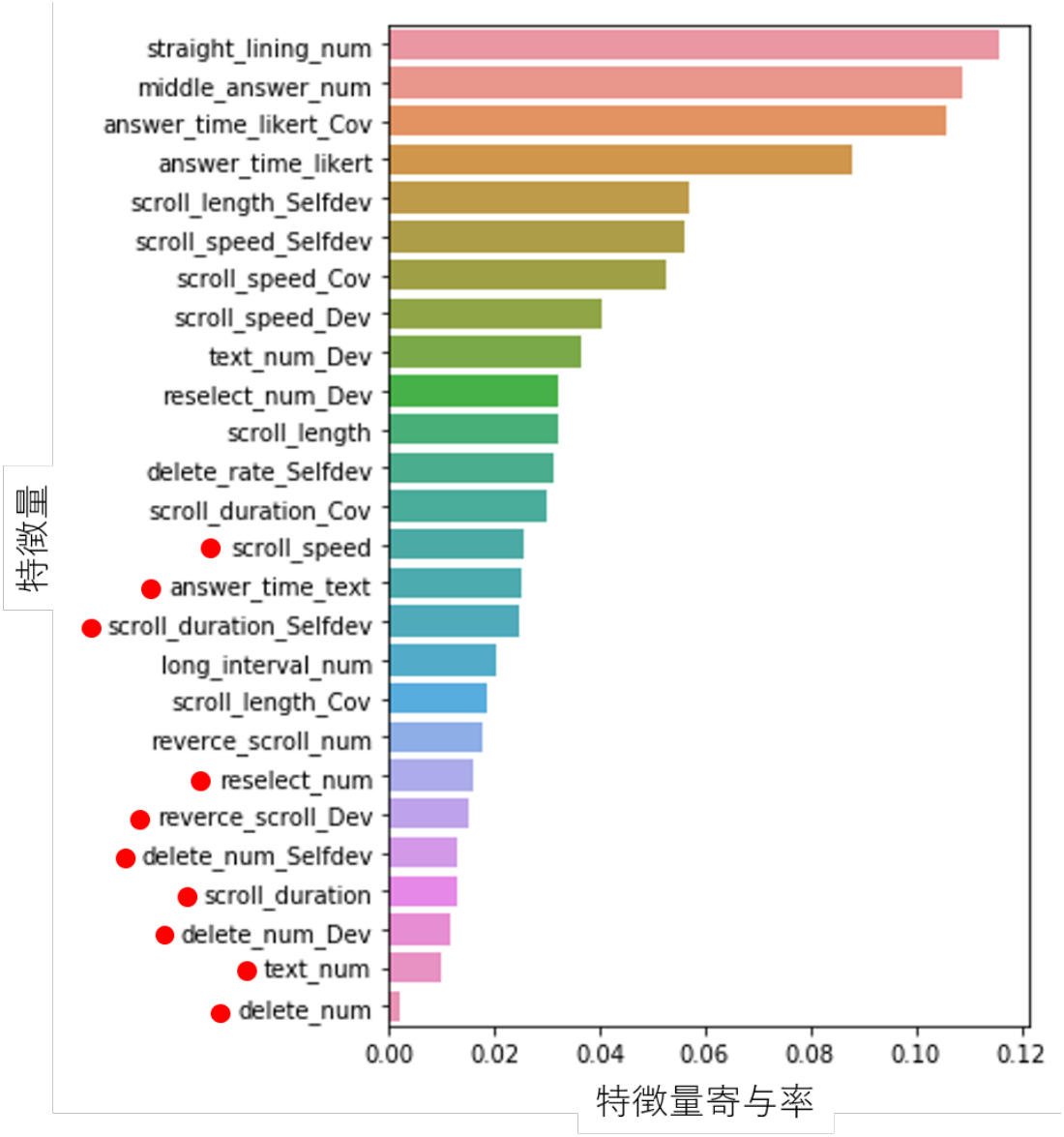


図 20 特徴量の寄与率（モデル 2）．モデル 3 では図中に赤丸で示す特徴量を削除した．

6.4 結果

3 つのモデルの評価結果を表 9 に示す．特に注目すべき Recall を含め，全ての指標でモデルを改良するにつれて精度が向上していることが確認できる．最終的なモデルの Recall は 85.9%であった．また，本分類問題において Precision が低いと，適切な回答を不適切であると誤って分類してしまう確率が高いということであるため，コストをかけて収集した適切なデータを無駄にしてしまうことに繋がる．また，偽陽性のサンプルを除くことはデータの代表性を不用意に損ねることになる．これらの観点から，Precision も Recall とともに大切な指標であると言える．Precision はモデル 1 からモデル 3 にかけて 2.3%向上した．これらの結果から，本研究ではスマートフォンにおける回答操作データを用いることで，尾崎ら [10] とほぼ同一の不適切回答の定義において，その検出精度を飛躍的に向上（55.6%→85.9%）可能であることを確認した．

表 9 機械学習モデルの評価結果

評価指標	モデル 1	モデル 2	モデル 3
Accuracy	0.844	0.850	0.862
Precision	0.841	0.848	0.864
Recall	**0.849**	**0.852**	**0.859**
F1 Score	0.845	0.850	0.862

6.5 モデルの頑健性

そこで，質問数と検出率の関係を検証するために，特徴量生成の対象ページ数（ベースライン算出に用いたページ数は除く）を 3 ページ（平均 17 問），9 ページ（90 問）とした場合，検出率はそれぞれ 79.7%，80.9%であった．6.4 節で述べた，全 17 ページ（128 問）を対象とした場合の 85.9%という結果も踏まえると，検証した範囲内においては，質問数が多いほどより高い精度で検出できる可能性が示唆された．一方で，17 問（3 ページ）程度の分量でも約 80%の検出率を維持することが確認された．

また，本稿で用いた機械学習モデルは決定木をベースとしたモデルであるため，例えば，不適切回答では適切回答よりも少ないはずの自由記述の文字数が多いなど，ミスリーディングな方向の外れ値は検出率を低下させる原因となる．そこで，このような外れ値に対する特徴選択モデルの頑健性を検証した．まず，全データを正例（不適切回答群）と負例（適切回答群）に分け，それぞれの群で各特徴量の平均値と四分位範囲を算出した．その平均値を比較して導かれた適切回答群と不適切回答群の大小関係を表 8 の「大小関係」欄に示す．次に，ミスリーディングな方向，つまり表 8 に示す大小関係に対して逆向きに四分位範囲を超える特徴量（以後,「外れ値」と呼ぶ）の数をサンプルごとに算出し，真陽性と偽陰性のサンプル数との関係を図 21 にプロットした．

横軸は外れ値の数を表す．棒グラフは真陽性と偽陰性それぞれのサンプル数，折れ線グラフは不適切回答の検出率を表す．折れ線グラフより，外れ値が多くなるに従って検出率が低下する傾向が確認できる．5 つ以上になると検出率は 50%以下になってしまう一方で，正例の約 80%を占める 3 つ以下の範囲で 86%以上の検出率を維持するモデルとなっており，本稿で用いた特徴量の外れ値に対して一定の頑健性を持つと考える．

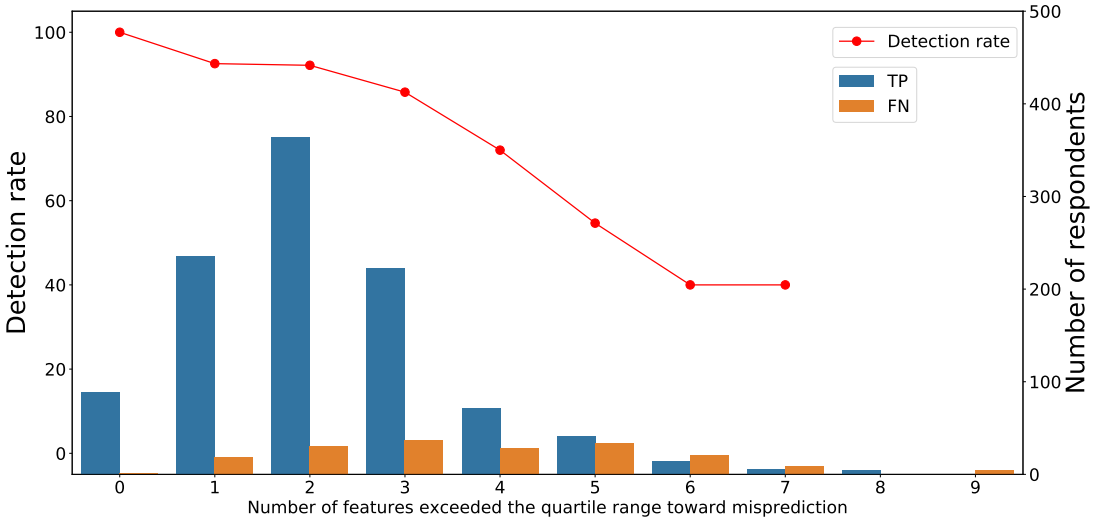


図 21 四分位範囲を誤検出の方向に越えた特徴量数と不適切回答の検出率（モデル 3）

7. 結論

本研究では，オンラインアンケートの信頼性を毀損する可能性がある Satisficing に基づく不適切回答の高精度な検出を目指した．これに向けて，スマートフォンの回答操作に着目したが，回答操作を記録できるアンケートシステムが存在しなかった．そこで，Satisficing 検出に寄与すると考えられる特徴量を提案し，回答操作が記録できる LimeSurvey 用プラグイン Operation Logger を開発した．これを用いてアンケート実験を行ない，回答結果とともに回答操作データを収集した．これらのデータから不適切回答検出に用いる特徴量を生成し，まずはその特徴量の有効性を統計検定によって確認した．1000 人に対して事前実験を行ない，考案した一部の特徴量で不適切回答群と適切回答群の間に有意差が確認できた．そこで次に，これらの特徴量を用いて機械学習による不適切回答の検出を行なうために追加でアンケートを実施し，事前実験のデータも含めて 5692 人のデータを収集した．正例 247 件，負例 4693 件という不均衡なデータであったため，負例をダウンサンプリングすることで正例と負例の比率が 1：1 のデータセットを作成し，機械学習モデルを作成した．絶対的な特徴量から回答者内・回答者間の偏差などの相対的な特徴量を追加し，さらに特徴量同士の相関によって特徴量を選択することで精度の向上を図った．結果として，85.9%という高い精度で不適切回答を検出できることが確認され，先行研究の精度を大幅に向上させることに成功した．また，高い検出率を実現するために必要な質問数の検証を行い，質問数が多い方が検出率が高くなる一方，17 問程度の質問数でも大幅に検出率が低下しないことを確認した．また，本稿で考案した特徴量の中では，スクロール速度やスクロール長など，スクロールに関する特徴量の寄与率が高かった．さらに，本稿で扱った特徴量全般について，回答者内および回答者間の偏差のような相対的な特徴量の方が，絶対的な特徴量よりも寄与する可能性が示唆された．

今後は、アンケートの冒頭から何ページ目もしくは何問目で不適切の判断ができるのかを検証する必要がある．さらに、その先の展望としては，本論文で述べた検出手法をもとに不適切回答を準リアルタイムに検出し，不適切な回答者に注意を促すよう介入するシステムの構築が考えられる．このとき，本研究ではアンケートのページごとに回答操作データをサーバに送信する設計となっているため，

送信スパンを短くする実装変更が必要となる．そのシステムを用いた実験を行ない，介入のない従来システムと比較することで，より高機能なシステムとしての評価ができると考える．さらに，アンケートページごとの各特徴量の推移なども特徴量に盛り込むなど，精度向上に寄与する新たな特徴量を模索する余地があると考える．

謝辞

本研究を進めるにあたり，安本慶一 教授には，研究全般に関し，ご指導・ご助言を賜りました．また，充実した研究環境の整備など，研究活動を手厚くご支援いただきました．感謝の意を表すとともに，心より厚く御礼申しあげます．

池田和司 教授には，ご多忙の中，論文審査委員を引き受けてくださった上で，副指導教官としてご助言をいただきました．感謝の意を表すとともに，心より厚く御礼申しあげます．

荒川豊 客員教授には，ご多忙の中，遠方から研究の方向性についての的確なご指導およびご助言をいただきました．特に，自分の研究を上手く他人に伝えるための技術について，丁寧に根気強くご教授くださり，非常に勉強になりました．感謝の意を表すとともに，心より厚く御礼申しあげます．

松田裕貴 助教には，本研究に関する的確なご助言や学会発表への同行，特許取得に向けた手続きなど，様々な面でご指導いただきお世話になりました．また，学生に近い立場で普段からコミュニケーションを取ってくださり，気軽に相談させて頂ける研究環境を提供してくださいました．感謝の意を表すとともに，心より厚く御礼申しあげます．

諏訪博彦 特任准教授には，本研究に関して，クラウドソーシングサービスの選定や，実験結果の統計分析に関するご助言・ご指導をいただきました．感謝の意を表すとともに，心より厚く御礼申しあげます．

金岡恵 事務補佐員，山内奈緒 事務補佐員，尾川恵理 事務補佐員には，学会や出張に関する事務処理を始め，研究生活の様々な場面でご支援いただきましたこと，謹んで感謝申し上げます．

大阪大学人間科学研究科の三浦麻子教授には，本研究のアンケート作成に関するご助言をいただきました．また，本研究で実施したアンケートの質問票に，同氏が公開されている質問票の一部を利用させていただきました．面識がないにも関わらず，快くミーティングの希望を受け入れてくださいましたこと，感謝の意を表すとともに，心より厚く御礼申しあげます．

九州大学システム情報科学研究院の Billy Dawnton 氏，Jihed Makhlouf 氏には，論文の添削をしていただき，的確なアドバイスをいただきました．心より感謝申

し上げます．

また，共に研究生活を過ごしたユビキタスコンピューティングシステム研究室の先輩，同輩，後輩には，公私ともにお世話になりました．心より感謝申しあげます．

最後に，今日まで学生生活を様々な面から支えてくださった家族に心より感謝申し上げます．

本研究の一部は，科研費（18H03233）および，JST さきがけ（JPMJPR2039）の助成で行われた．

参考文献

[1] 総務省統計局. 国勢調査のあゆみ. https://www.stat.go.jp/data/kokusei/2015/kouhou/ayumi.html. (Accessed on 2021/01/25).

[2] M. R. Maniaci and R. D. Rogge. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, Vol. 48, pp. 61–83, 2014.

[3] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, Vol. 63, No. 2, pp. 129–138, 1956.

[4] J. A. Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, Vol. 5, No. 3, pp. 213–236, 1991.

[5] D. M. Oppenheimer, T. Meyvis, and N. Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, Vol. 45, No. 4, pp. 867 – 872, 2009.

[6] A. Miura and T. Kobayashi. Survey satisficing inflates stereotypical responses in online experiment: The case of immigration study. *Frontiers in Psychology*, Vol. 7, p. 1563, 2016.

[7] 三浦麻子, 小林哲郎. オンライン調査における努力の最小限化が回答行動に及ぼす影響. 行動計量学, Vol. 45, No. 1, pp. 1–11, 2018.

[8] D. J. Hauser and N. Schwarz. It ' s a trap! instructional manipulation checks prompt systematic thinking on “ tricky ” tasks. *SAGE Open*, Vol. 5, No. 2, p. 2158244015584617, 2015.

[9] P. Weiping, M. Arthur, T. Kaylynn, and Y. Chuan. Attention please: Your attention check questions in survey studies can be automatically answered. In *Proceedings of The Web Conference 2020*, WWW '20, p. 1182–1193, New York, NY, USA, 2020. Association for Computing Machinery.

[10] 尾崎幸謙, 鈴木貴士. 機械学習による不適切回答者の予測. 行動計量学, Vol. 46, No. 2, pp. 39–52, 2019.

[11] R. Tourangeau, H. Sun, T. Yan, A. Maitland, G. Rivero, and D. Williams. Web surveys by smartphones and tablets: Effects on data quality. *Social Science Computer Review*, Vol. 36, No. 5, pp. 542–556, 2018.

[12] 後上正樹, 松田裕貴, 荒川豊, 安本慶一. オンラインアンケートの回答信頼性検証に向けた回答時画面操作ログ取得システム. 情報処理学会研究報告, Vol. 2020-HCI-186, No. 35, pp. 1–7, 2020.

[13] 三浦麻子, 小林哲郎. オンライン調査モニタの satisfice に関する実験的研究. 社会心理学研究, Vol. 31, No. 1, pp. 1–12, 2015.

[14] D. R. Mandel. Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General*, Vol. 143, No. 3, pp. 1185–1198, 2014.

[15] M. Revilla and C. Ochoa. What are the links in a web survey among response time, quality, and auto-evaluation of the efforts done? *Social Science Computer Review*, Vol. 33, No. 1, pp. 97–114, 2015.

[16] 三浦麻子, 小林哲郎. オンライン調査における努力の最小限化（satisfice）を検出する技法：大学生サンプルを用いた検討. 社会心理学研究, Vol. advpub, , 2016.

[17] 増田真也, 坂上貴之, 森井真広. 調査回答の質の向上のための方法の比較. 心理学研究, Vol. 90, No. 5, pp. 463–472, 2019.

[18] 深井裕二, 河合洋明. Moodle アンケートに対応した satisfice 回答の適応的除去システムの開発. 工学教育, Vol. 65, No. 3, pp. 60–65, 2017.

[19] P. Lugtig and V. Toepoel. The use of pcs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error. *Social Science Computer Review*, Vol. 34, No. 1, pp. 78–94, 2016.

[20] NTT コムオンライン・マーケティング・リサーチ株式会社. 回答結果の品質 : 回答結果の品質向上のための取り組み. http://research.nttcoms.com/service/qpolicy4.html. (Accessed on 2021/01/25).

[21] Y. Kim, J. Dykema, J. Stevenson, P. Black, and D. P. Moberg. Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Soc. Sci. Comput. Rev.*, Vol. 37, No. 2, pp. 214–233, 2019.

[22] H. Baumgartner and J.-B. E.M. Steenkamp. Response styles in marketing research: A cross-national investigation. *J. Mark. Res.*, Vol. 38, No. 2, pp. 143–156, 2001.

[23] N. A. Bowling, A. M. Gibson, J. W. Houpt, and C. K. Brower. Will the questions ever end? person-level increases in careless responding during questionnaire completion. *Organ. Res. Methods*, p. 1094428120947794, 2020.

[24] Y. Hirabe, H. Suwa, Y. Arakawa, and K. Yasumoto. Touchanalyzer: A system for analyzing user ' s touch behavior on a smartphone. *International Journal of Computer Science and Mobile Computing*, Vol. 7, pp. 25–38, 2018.

[25] Contentsquare. Clicktale. https://www.clicktale.com/. (Accessed on 2021/01/25).

[26] Google, LLC. Google forms. https://www.google.com/forms/about/. (Accessed on 2021/01/25).

[27] SurveyMonkey. Surveymonkey. https://www.surveymonkey.com/.

[28] 三浦麻子, 小林哲郎. Supplemental materials for "satisficing" studies by miura, a. and kobayashi, t. https://osf.io/6gu3q/. (Accessed on 2021/01/25).

[29] K. Guolin, M. Qi, F. Thomas, W. Taifeng, C. Wei, M. Weidong, Y. Qiwei, and L. Tie-Yan. Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing*

Systems, NIPS'17, p. 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc.

[30] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2623—-2631, New York, NY, USA, 2019. Association for Computing Machinery.

研究業績

学術論文

1. Masaki Gogami, Yuki Matsuda, Yutaka Arakawa, Keiichi Yasumoto :“Detection of Careless Responses in Online Surveys Using Answering Behavior on Smartphone”, IEEE Access (Accepted but not published yet).

国内会議

1. 後上 正樹, 谷 優里，松田 裕貴, 荒川 豊, 安本 慶一 :“オンラインアンケートの回答信頼性に影響する指標の調査検討- アンケート形式とスマートフォン操作状況の観点から –”, 情報処理学会関西支部大会, 大阪府, 2019 年 9 月.

2. 後上 正樹, 松田 裕貴, 荒川 豊, 安本 慶一 :“オンラインアンケートの回答信頼性検証に向けた回答時画面操作ログ取得システム”, IPSJ SIG-HCI, 沖縄県, 2020 年 1 月.

3. 後上 正樹, 松田 裕貴, 荒川 豊, 安本 慶一 :“オンラインアンケートにおける不適切回答自動検出に向けた回答操作ログ分析”, 第 13 回データ工学と情報マネジメントに関するフォーラム (DEIM 2021), オンライン, 2021 年 3 月.

4. 後上 正樹, 松田 裕貴, 荒川 豊, 安本 慶一 :“オンラインアンケート回答時のスマートフォン画面操作状況に基づく不適切回答検出”, 第 25 回一般社団法人情報処理学会シンポジウム インタラクション 2021, オンライン, 2021 年 3 月.

受賞

1. DEIM 学生プレゼンテーション賞 : “オンラインアンケートにおける不適切回答自動検出に向けた回答操作ログ分析”, 第 13 回データ工学と情報マネジメントに関するフォーラム (DEIM 2021).

特許

1. 後上 正樹, 松田 裕貴, 荒川 豊, 安本 慶一 :“オンラインアンケートの不適切回答検出システム”, 2020 年 12 月.