日本データベース学会和文論文誌
Vol. 20-J, Article No. 9, 2022年 3月

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の統計分析

後上正樹[1]　松田裕貴[2]　荒川豊[3]　安本慶一[4]

オンラインアンケートにおいて，なるべく楽に早くアンケートタスクを完了しようとする「Satisficing（努力の最小限化）」という態度が調査結果の信頼性を低下させる問題がある．この問題に対し，回答時間が明らかに短い回答を除外するなどの措置が考案されてきたが，回答時間だけでは完全に除外することができない．そこで我々は，回答時間だけでなく，回答中の画面操作を観測することで，より高精度にSatisficingが検出可能になるのではないかという仮説を立てた．世界中で利用されているオープンソースアンケートシステムLimeSurveyの拡張機能として，回答中の画面操作ログ（回答操作ログ）を記録できるプラグインを開発し，本システムを用いてクラウドソーシング上でアンケート実験を行った．実験は，回答者は各自のスマートフォンを用いて，リッカート形式と自由記述形式の質問から成るアンケートに回答してもらうものであり，アンケートの途中にはSatisficingを検出するために考案されている既存の質問を挿入した．1000人に対して回答を依頼し，収集した回答内容および回答操作ログから，Satisficingを表現する可能性が考えられる特徴量を生成し，特徴量ごとにSatisficingである群とSatisficingでない群の母平均の差の検定を行った．結果として，既存の特徴量であるテキストの文字数，連続同一回答数，中間回答数に加え，開発したプラグインでのみ取得可能なスクロール速度，選択肢の変更回数，リッカート形式の回答時間などに統計的有意差があることを明らかにした．また，スクロール速度やテキストの変更回数に関しては，絶対値では差がない一方で，回答者ごとのベースラインとの差では有意な差が確認できた．

1　はじめに

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る．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 [1] 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 [2] は，人間の認知的資源には限りがあることによって生じる，アンケート調査において回答要求に対する努力を最小化しようとする傾向を努力の最小限化（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 [3] はIMC（Instructional Manipulation Check），Maniaciら [1] 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 [4–6]．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．そこで尾崎ら [7] は，従来のような質問の挿入を不要とする機械学習による不適切回答の検出を試みた．様々な機械学習手法を用いた結果，不適切回答の検出率が最も高かったモデルで65.5%と報告されている．

本研究ではスマートフォンの回答操作ログを記録し，ログデータから生成する操作特徴量に基づくSatisficing検出手法により検出率の向上を目指す．この目標に対して本稿では，オープンソースアンケートシステムLimeSurveyの拡張機能として，回答操作ログを記録するプラグインを開発し [8]，1000人を対象にアンケート実験を行った．収集した回答操作ログからSatisficingを表現すると考えられる特徴量を生成し，Satisficingである群（以下，Satisficing群）とSatisficingでない群（以下，正常群）それぞれから得られる各特徴量について母平均の差の検定を行なった．結果として，既存の特徴量であるテキストの文字数，連続同一回答数，中間回答数に加え，開発したプラグインでのみ取得可能なスクロール速度，選択肢の変更回数，リッカート形式の回答時間などに統計的有意差があることを明らかにした．また，スクロール速度やテキストの変更回数に関しては，絶対値では差がない一方で，回答者ごとのベースラインとの差では有意な差が確認された．

本稿の構成は次の通りである．2章で関連研究について述べる．3章でSatisficingと関係があると思われる回答操作ログについて述べる．4章では回答操作ログデータを記録するアンケートシステムについて述べる．5章でそのシステムを用いたアンケート実験について説明し，6章で分析結果と考察を述べる．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 [3] 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，

[1] 非会員　奈良先端科学技術大学院大学
gogami.masaki.gg8@is.naist.jp
[2] 非会員　奈良先端科学技術大学院大学／JSTさきがけ
yukimat@is.naist.jp
[3] 非会員　九州大学
arakawa@ait.kyushu-u.ac.jp
[4] 非会員　奈良先端科学技術大学院大学
yasumoto@is.naist.jp

DBSJ

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 [9] 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 [10, 11].

この問題に対して，Maniaciら [1] は，ARS（The 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 [12] 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確定的な絞り込みは実現できなかったと述べられている．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．さらに，Peiら [13] 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，これらのスクリーニング質問を用いた検出手法は将来的に信頼性が脅かされる可能性がある．

そこで，スクリーニング質問を追加することなくSatisficing検出を実現するために，尾崎ら [7] はPCによって回答された結果について，機械学習を用いてSatisficingの検出を試みた．説明変数には，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値など，回答結果から得られる情報を用いて，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65.5%という結果を報告している．また，我々はこれまでの研究で，スマートフォンの回答操作ログの機械学習に基づくSatisficing検出モデルを構築しており，DQSとARSの両方でSatisficingと判定されたサンプルを85.9%の再現率で検出可能であることを確認している [14, 15].

表1　Satisficing検出のために記録する特徴量

ログデータの種類	単位	質問形式	独自追加
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	○
テキストの変更回数	回	自由記述	○
テキストボックスの再フォーカス回数	回	自由記述	○
スクロール長	px	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	○
アンケート全体の回答時間	s	全体	–
最大連続同一回答数	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	–

本研究では機械学習モデルの特徴量として，スマートフォンの回答操作特徴を用いることで高精度にSatisficingが検出できるモデルの構築を最終的な目標としている．そのために本稿では，特徴量ごとにSatisficing群と正常群の間で統計的な差があるかどうか検証した．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 [16]． Rogerら [17] 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を検出するための特徴量

Satisficingを検出するために有効であると考える特徴量を表1に示す．表中の「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．また，「全体」の項目は質問形式に関係なくアンケート全体に関する特徴量である．「独自追加」欄では，既存のアンケートシステムで記録可能であった特徴量については「–」印，我々が新たに記録する特徴量については「○」印で記載している．スクロールや回答の変更に関する特徴量は今までのアンケートシステムでは記録できなかった．他の既存の特徴量についても，今まで報告されているものを概ねカバーする．

「回答時間」は，これまでの研究でも不適切回答検出のための特徴量として用いられてきた．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 [18, 19]．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．しかしながら，依然として回答時間はSatisficingに強く関係する特徴量であると考える．本稿で

特徴量として扱う回答時間は，「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する．

「選択肢の変更回数」，「テキストの変更回数」，「テキストボックスの再フォーカス回数」は，正確な回答を試みている状態が表れると考えられる．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，これらの特徴量もSatisficingに関連すると考える．

「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する．また，「スクロール速度」は「スクロール長」を一回のスクロール操作にかかる時間で除算した値と定義する．これらは質問間の移動という操作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，100 px以上の逆向きのスクロールを1回と定義した．100 px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必要とする移動量であることに基づく．これは，アンケートの回答中に前の質問の回答を変更したくなったり，ページ冒頭の質問文を読み直す際の操作を表していると考えられる．このような操作は丁寧な回答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高いと考えられる．

「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る．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，この特徴量が大きい場合はSatisficingが発現している可能性が高いと考えられる．

「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である．Satisficingがない状態でも同一回答になる場合はあるが，Satisficingが発現している回答者はその数が異常に多くなる場合がある [20]．

「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がない状態でも中間回答になる場合はある．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コストをかけ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 [21]．そのため，中間回答数の多さがSatisficingの検出に寄与すると考えられる．

「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に関連していると考えられ，文字数少ない方がSatisficing傾向が強いであろうと考える．

また，リッカートの回答時間，スクロール長，スクロール速度について，分散を捉える変動係数を特徴量に追加した．さらに，これまでに記載した特徴量は絶対的な値であるが，被験者間および被験者内での相対的な値も有効な特徴量になり得ると考えた．そこで，テキスト文字数，テキストの変更回数，選択肢の変更回数，逆スクロール回数に関しては被験者間の偏差を，テキストの変更回数，スクロール長，スクロール速度に関しては被験者ごとのベースラインとの差を算出し，特徴量に追加した．被験者ごとのベースラインは，アンケート内の4，5ページ目において計測した．

図1　回答操作記録アンケートシステムの概観

4　回答操作記録アンケートシステム

4.1　システム構成

我々は，2章で述べた特徴量を記録するシステム構成を設計した [8]．その際，システムの普及性を考慮した結果，新規にアンケートシステムを作るのではなく，既存のシステムを拡張する方法が最も合理的であると判断した．ClickTale等のWebサイトの顧客体験改善のためにユーザのページ内行動を可視化するサービスがあるが，汎用性が高い反面，アンケートに特化した特徴量の記録はできない．そこで，オープンソースのオンラインアンケートシステムであるLimeSurveyに着目した．LimeSurveyの標準機能で記録できるデータは，最終的な回答内容，回答開始時刻，回答終了時刻，ページごとの回答時間である．一方，Google FormやSurvey Monkey等とは異なり，Javascript（以下js）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に示す回答操作ログを取得するLimeSurveyプラグインを独自に開発し，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．本プラグインの導入方法は，これら3つのファイルをLimeSurveyをホスティングするサーバに配置し，質問設定画面のヘルプ文章欄にてjsファイルの読み込み設定をするのみである．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．

4.2　提案システムで記録されるデータ

開発したプラグインで記録されるデータの詳細について述べる．データを取得するイベントはタッチイベント，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

表2　提案システムの動作確認済みモバイル端末環境

	Chrome	Safari	Firefox	Opera
iOS	✓	✓	✓	✓
Android	✓	N/A	✓	✓

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」，「スクロール速度」，「逆スクロール回数」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，レコード生成理由を取得し，データベースに格納する．これにより，自由記述形式の「質問単位の回答時間」と「テキストの変更」が検出できる．テキストの記録単位の程度を検討した結果，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．なお，本システムは表2の「✓」印で示す環境で所望の動作を確認した．

5　クラウドソーシングを用いたアンケート実験

４章で述べた回答操作記録アンケートシステムを用いて実施したアンケート実験について説明する．なお，本研究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）．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で行なった．クラウドソーシングサービスはYahoo!クラウドソーシングを用いて，被験者数は1000人とした．被験者の募集に際して，ワーカーをプラットフォーム上の指標（ブラックリストなど）でフィルタリングせず，全てのワーカーが1回のみ回答できるものとした．回答者には報酬として5円相当のポイントを付与した．実験で用いたSatisficing指標とアンケート設計について以下で述べる．

5.1　Satisficing指標

本実験では，回答者単位でSatisficingか否かのラベルを与える指標として既存研究2種類用いた．1つ目がDQSであり，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．今回は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．2つ目がARS（Attentive Responding Scale）であり，これにはさらに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 [1]. 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．三浦ら [5] が用いたIMCに関しては，今回用いる質問票が特別な教示を必要としないため，適さないSatisficing指標であると判断して使用しなかった．

図2　質問項目の概略

5.2　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2に示す．この質問票は三浦ら [12] が公開している質問票 [22] 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．

1〜3ページ目は基本情報を取得するための質問である．4ページ目は被験者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ページ目は被験者ごとのスクロール速度およびスクロール長のベースラインを測定するための質問である．質問内容としてはいずれかの選択肢を選択するように指示するもので，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測定した．これは，後述する特徴量である被験者内の相対的スクロール速度および相対的スクロール長を算出するために測定した．6ページ目は被験者ごとのDelete率のベースラインを測定するための質問である．指示文章の入力時のDelete数を文字数で除算したものを各被験者のDelete率のベースラインとした．これは後述する特徴量である被験者内の相対的Delete率を算出するために測定した．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．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．

6　実験結果および考察

被験者1000人のうち，回答操作ログデータの使用に同意した817人のデータを分析対象とした．まずは回答結果を基に，DQS・ARSそれぞれでSatisficingラベルを付与した．DQSに関

表3　DQS，ARSの該当者数および割合

Satisficing ラベル	カテゴリ	該当者数［人］	割合［%］
no DQS	遵守	732	90
DQS	違反	85	10
no ARS	遵守	672	82
OR	違反	128	16
AND	違反	17	2

表4　正常群に対するSatisficing群の母平均の大小と検定結果

特徴量	母平均の大小[a]	有意差[b] DQS	ARS
リッカートの回答時間	Low	∗	∗
自由記述の回答時間	Low	–	–
選択肢の変更回数	Low	∗	∗
テキストの変更回数	Low	–	–
テキストボックスの再フォーカス回数	Low	–	–
スクロール長	Low	–	–
スクロール速度	High	∗	∗
逆スクロール回数	Low	–	–
非操作時間が長すぎる回数	Low	–	–
最大連続同一回答数	High	∗	∗
中間回答数	High	∗	∗
文字数	Low	∗	∗
リッカートの回答時間の変動係数	High	–	∗
スクロール長の変動係数	Low	–	–
スクロール速度の変動係数	High	–	–
文字数の偏差	High	–	–
テキストの変更回数の偏差	High	∗	–
選択肢の変更回数の偏差	High	∗	∗
逆スクロール回数の偏差	High	–	–
テキストの変更回数の自己BLとの差	High	∗	–
スクロール長の自己BLとの差	High	∗	∗
スクロール速度の自己BLとの差	High	∗	∗

a) 正常群の母平均に対しSatisficing群の母平均が大きい（High）か小さい（Low）かを示す．
b) 「–」印は有意差がないこと，「∗」印は有意差があることを表す．

(a) DQS　(b) ARS

図3　選択肢の変更回数

(a) DQS　(b) ARS

図4　スクロール速度

(a) DQS　(b) ARS

図5　リッカートの回答時間

しては，DQS 3問のうち全問指示に従った場合は「no DQS」，1問以上指示に従わなかった場合は「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「no ARS」，どちらか1つがSatisficingであるサンプルを「OR」，どちらもSatisficingであるサンプルを「AND」というラベルに割り当てた．各ラベルに対応するサンプル数を表3に示す．次に，Satisficing群と正常群の母平均の差を有意水準5%とするウェルチのt検定によって検定した．ここで，Satisficing群と正常群の割り当ては表3の「カテゴリ」欄に示す通りとし，ARSに関してはSatisficing群を「OR」と「AND」の和集合とした．

全特徴量の検定結果を表4に示す．「母平均の大小」欄では，正常群とSatisficing群の母平均を比較したとき，Satisficing群の方が大きい（High）か小さい（Low）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．

また，有意差が認められた特徴量の一部について，正常群とSatisficing群との比較を図3，4，5，6に示す．図中の「∗」印は有意水準5%で有意差があったことを表している．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6(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と考えられる．

表4を見ると，各種変動係数は統計的な差がほとんど見られ

(a) DQS　(b) ARS

図6　文字数

(a) DQS　(b) ARS

図7　逆スクロール回数

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，偏差に関しては，絶対値では有意差があった文字数について有意差は見られなかった．一方で，テキストの変更回数は絶対値では有意な差が見られなかったものの，偏差ではDQSに限って有意差が見られた．各種ベースラインとの差に関しては，概ね有意な差が認められた．

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7，8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7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正常群に属している．これより，統計的には差があったとは言えないものの，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9に示すベースラインとの差を取った特徴量ではDQS，ARS共に有意な差が認められた．テキストの変更回数に関しても，表4を参照するとベースラインとの差をとった特徴量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8　スクロール長

(a) DQS　(b) ARS

図9　スクロール長の自己ベースラインとの差

7　まとめと今後の展望

本稿では，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の検出に向けて，オープンソースアンケートシステムLimeSurvey向けに開発したプラグインを用いたアンケート実験の結果について述べた．本プラグイン導入することで，これまでの既存アンケートシステムでは記録できなかった回答操作ログデータを記録することが可能となった．これらのデータから生成した各種特徴量について，正常群とSatisficing群の母平均の差の検定を行なった結果，スクロールや回答の変更等にSatisficingが表出するという仮説の妥当性を示す統計的な有意差が確認された．また，スクロール速度やテキストの変更回数に関しては，絶対値では差がない一方で，回答者ごとのベースラインとの差を取った値では有意な差が確認できた．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．

今後は実験を継続してデータ数を増やし，上記のような知見を踏まえてSatisficingを検出する機械学習モデルの開発に取り組む．また，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．さらに，質問の順序による回答操作への影響などについても調査する必要がある．

謝辞

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．また，大阪大学人間科学研究科の三浦麻子教授が公開している質問票を，本稿のアンケート実験のベース質問票として用いさせて頂いた．さらに，Satisficing検出質問を含むアンケート設計についてご教示頂いた．ここに同氏に対して謝意を表する．

参考文献

[1] Michael R. Maniaci, Ronald D. Rogge, “Caring about carelessness: Participant inattention and its effects on research,” Journal of Research in Personality, Vol. 48, pp. 61–83, 2014.
[2] Herbert A. Simon, “Rational Choice and the Structure of the Environment,” Psychological Review, Vol. 63, No. 2, pp.129–138, 1956.
[3] Daniel M. Oppenheimer, Tom Meyvis, Nicolas Davidenko, “Instructional manipulation checks: Detecting satisficing to increase statistical power,” Journal of Experimental Social Psychology, Vol. 45, No. 4, pp. 867–872, 2009.
[4] Asako Miura, Tetsuro Kobayashi, “Survey satisficing inflates stereotypical responses in online experiment: The case of immigration study,” Frontiers in Psychology, Vol. 7, Article 1563, pp. 1–10, 2016.
[5] 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.
[6] David J. Hauser, Norbert Schwarz, “It’s a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks,” SAGE Open, Vol. 5, No. 2, pp. 1–6, 2015.
[7] 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.
[8] 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. 1–6, 2020.
[9] 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.
[10] David R. Mandel, “Do framing effects reveal irrational choice?,” Journal of Experimental Psychology: General, Vol. 143, No. 3, pp. 1185–1198, 2014.
[11] Melanie Revilla, Carlos 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.
[12] 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. 32, No. 2, pp. 123–132, 2016.
[13] Weiping Pei, Arthur Mayer, Kaylynn Tu, and Chuan Yue, “Attention Please: Your Attention Check Questions in Survey Studies Can Be Automatically Answered,” Proceedings of The Web Conference 2020, pp. 1182–1193, 2020.
[14] 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,” 情報処理学会インタラクション2021, pp. 11–20, 2021.
[15] Masaki Gogami, Yuki Matsuda, Yutaka Arakawa and Keiichi Yasumoto, “Detection of Careless Responses in Online Surveys Using Answering Behavior on Smartphone,” IEEE Access, Vol.9, pp.53205-53218, 2021.
[16] Peter Lugtig, Vera 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.
[17] Roger Tourangeau, Hanyu Sun, Ting Yan, Aaron Maitland, Gonzalo Rivero, Douglas Williams, “Web Surveys by Smartphones and Tablets: Effects on Data Quality,” Social Science Computer Review, Vol. 36, No. 5, pp. 542–556, 2018.
[18] NTTコムオンライン・マーケティング・ソリューション株式会社, “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（2021年5月14日閲覧）.
[19] 株式会社マクロミル, “モニタの品質管理ポリシー：品質を高めるシステム機能,” https://www.macromill.com/advantage/monitor_policy.html（2022年1月29日閲覧）.
[20] Yujin Kim, Jennifer Dykema, John Stevenson, Penny Black, D. Paul Moberg, “Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys,” Social Science Computer Review, Vol. 37, No. 2, pp. 214–233, 2019.
[21] Hans Baumgartner, Jan-Benedict E.M. Steenkamp, “Response Styles in Marketing Research: A Cross-National Investigation,” Journal of Marketing Research, Vol. 38, No. 2, pp. 143–156, 2001.
[22] Asako Miura, Tetsuro Kobayashi, “Supplemental materials for "Satisficing" studies by Miura, A. and Kobayashi, T.,” https://osf.io/6gu3q/（2022年1月29日閲覧）.