Vol.2021-SPT-41 No.42
2021/3/2

オンラインユーザ調査における参加者の注意力と回答内容の分析

松浦 天我[1,a)]　長谷川 彩子[3]　秋山 満昭[3]　森 達哉[1,2,4]

概要：人間の認識や行動の理解を目的とする学術研究において，より効率的な回答の収集が期待できる手法として，クラウドソーシングサービスを用いたオンラインユーザ調査が広く行われている．オンラインユーザ調査では，多数の作業をこなすクラウドワーカーによる不誠実・不注意な回答や，ボットによる回答が含まれるリスクがあるため，このような不良回答を除去するための手法がいくつか提案されている．しかし，セキュリティ研究，特に，参加者の注意深さが結果に影響すると考えられるフィッシング研究においては，安直な不良回答除去の適用により，本来研究対象とすべき参加者の除去に繋がる可能性がある．そこで本研究では，「**既存の不良回答除去手法を適用することにより，セキュリティ研究のアンケート結果に偏りは生じ得るか？**」という研究的問いに取り組む．300 名の参加者を対象としてフィッシングメールの特定を題材としたユーザ調査を実施した結果，不誠実な参加者が全体の 40%を占めること，回答完了時間および自由記述式質問による不良回答除去には一定の効果があること，フィッシングに関するユーザ調査において安直な IMC の適用は研究結果に偏りを生じさせる原因となり得ることを明らかにした．

キーワード：オンラインユーザ調査，クラウドソーシングサービス，フィッシング，セキュリティ行動

An analysis of participants' attention and response in online user surveys

Abstract: In the academic research projects that aim to understand human perception and behavior, online user studies using crowdsourcing services have been widely used as a means to collect many responses efficiently. Since there is a risk that online user surveys may include dishonest or careless responses by crowd workers who perform a large number of tasks, or responses by bots, several methods have been proposed to eliminate such faulty responses. However, in security research, especially in phishing research where the attentiveness of the participants is considered to affect the results, the removal of faulty responses may lead to the removal of participants who should be included in the research. In this study, we address the following research question: "*Does the adoption of existing faulty answer removal methods bias the results of security research questionnaires?* An online user study of 300 participants on the subject of phishing email identification revealed that dishonest participants accounted for 40% of the total participants, that faulty answer removal using response completion time and open-ended questions was effective in extracting meaningful responses, and that in a user study on phishing, a careless adoption of the instructional manipulation check (IMC) can lead to biased results.

Keywords: Online User Study, Crowdsourcing, Phishing, Security Behavior

1. はじめに

人間の認識や行動の理解を目的とする学術研究においてはしばしばユーザ調査が実施され，その実施環境は実験室環境からオンライン環境に移行している．COVID-19 の影響もあり，ユーザ調査のオンライン化はさらに加速するも

[1] 早稲田大学 (Waseda University)
[2] 情報通信研究機構 (NICT)
[3] NTT セキュアプラットフォーム研究所 (NTT Secure Platform Laboratories)
[4] 理化学研究所 革新知能統合研究センター (RIKEN AIP)
[a)] tenga1012@nsl.cs.waseda.ac.jp

のと考えられる．

オンラインでのユーザ調査には広く参加者を募集できるなどの様々なメリットがある一方，いくつかのデメリットも知られている．主要なデメリットは，人間による不誠実および・不注意な回答，ボットによる自動回答といった「**不良回答**」が含まれることである．

これまで様々な不良回答除去手法が社会心理学分野を中心に提案され，実際の学術研究において頻繁に実装されている．中でも，注意力の高くない参加者を除去する手法がよく用いられるが，結果として特定の特徴をもつ参加者が除去されデモグラフィーの偏りが生じてしまうことが明らかになっている [8], [16]．我々は，セキュリティ研究，特に，参加者の注意深さが結果に影響すると考えられるフィッシング研究においては，注意力の高くない参加者を除去することは，本来研究対象とすべき参加者を除去することに繋がる可能性があると考えた．

そこで本研究では，以下の研究的問い (Research Question: RQ) に取り組む．

RQ: 既存の不良回答除去手法を適用することにより，セキュリティ研究のアンケート結果に偏りは生じ得るか？

この RQ に対し，本研究は不良回答除去手法を用いて参加者を誠実さや注意力の程度によりグループに分類し，各グループのセキュリティ知識/行動やフィッシングメール特定パフォーマンスを比較する．

クラウドソーシングサービスを用いて 300 人のアメリカ在住者を対象にアンケート調査を実施した結果，不誠実な参加者が全体の 40%を占めることが明らかになった．また，参加者の注意力の程度で回答除去を実施した際に，有効となる参加者の性別・学歴などに偏りが生じることを確認した．さらに，注意力が中程度の参加者と高程度の参加者では，フィッシングメールに対する判断の傾向が異なることが明らかにした．本研究で得られた結果をもとに，セキュリティ研究における不良回答除去手法の適切な実施方法について議論する．

本論文の貢献を以下に示す．

- 既存の不良回答除去手法を適用することにより，セキュリティに関するアンケート調査の結果に偏りが生じ得ることを実証した初の研究である．
- 特にフィッシングサイトに対するユーザ行動の研究においては，参加者の注意力に基づく不良回答除去手法の適用は不適切であることを示した．

本論文の構成は以下の通りである．2 章で研究背景をまとめる．次に 3 章にて調査手法を，4 章にて調査の結果を示す．5 章では，結果をもとに不良回答除去手法の有効性や課題について議論した後，本研究の制約事項，将来の研究課題に関して述べる．最後に 6 章で本研究についてまとめる．

2. 研究背景

本章では，クラウドソーシングサービスを利用したオンラインユーザ調査における不良回答の問題と，関連する研究をまとめる．

Amazon Mechanical Turk（MTurk）[1] を始めとしたクラウドソーシングサービスが一般のインターネットユーザに広がるにつれ，学術研究においてもクラウドソーシングサービスを利用したオンラインユーザ調査が頻繁に実施されるようになった．オンラインユーザ調査には，多様な参加者を短期間で大量に募集できるというメリットがあり，研究者にとって利便性が高い．その一方で，オンラインユーザ調査では，人間による不誠実・不注意な回答，ボットによる自動回答といった不良回答が含まれることが問題になっている．特に MTurk においては，不良回答が 2018 年頃から急増しているとされる [10]．ユーザ調査における不良回答の発生には，参加者に金銭報酬が授与されることに加え，匿名（厳密には仮名）回答形式であることやオンラインでは調査者による監視が実施されないことなどが影響しているとされる [6]．

オンラインユーザ調査における不良回答の除去に向けて，これまでに様々な種類の不良回答除去手法が社会心理学分野を中心に提案され，その効果が検証されてきた．例えば，Yarrish ら [17] や Buchanan ら [4] は，CAPTCHA，回答時間，自由記述回答，注意力テスト，Instructional manipulation check（IMC）[11] などによる不良回答除去の効果を調査し，その有効性を確認した．これら 5 種類の不良回答除去手法の詳細は 3.2 節で述べる．不良回答除去手法には一定の効果が見込まれることから，多くの社会心理学研究者が不良回答除去手法の実装を推奨している **?**．我々が実施した予備調査により，セキュリティ研究においても，自由記述回答，注意力テスト，IMC による不良回答除去が実際に行われていることが確認された．

しかしながら，不良回答手法の実装により結果として特定の特徴をもつ参加者が除去されデモグラフィーの偏りが生じてしまう可能性があることから，その実装には注意が必要である [8], [16]．我々は，セキュリティ研究，特に，参加者の注意深さが結果に影響すると考えられるフィッシング研究においては，注意力の高くない参加者を除去する不良回答除去手法の実装は，本来研究対象とすべき参加者を除去することに繋がる可能性があると考えた．本研究では，既存の不良回答除去手法を適用することにより，セキュリティ研究のアンケート結果に偏りが生じ得るかを調査する．

3. 調査手法

本研究では，不良回答除去手法の実装がセキュリティ研

究のアンケート結果に偏りを生じさせるかどうかを調査するために，クラウドソーシングサービスを用いてオンラインアンケートを実施する．以下では，実験に用いた質問紙，不良回答除去手法，および参加者の募集方法について説明する．

3.1 質問紙

我々が作成した質問紙は計 53 問からなり，参加者のデモグラフィーを問う質問，セキュリティ知識・行動等を問う質問，フィッシングメール特定タスクの 3 パートに分けられる．複数の不良回答除去手法が質問紙の中に挿入された．質問紙の実装にはオンラインアンケート作成サービス Qualtrics [15] を利用し，質問紙は英語で記述した．

パート 1：デモグラフィーを問う質問

パート 1 では，参加者の年齢，性別，最終学歴，IT 職歴の有無，利用デバイスおよび利用時間を問う質問を設けた．先行研究 [8] において，IMC の実装によって若い人，男性，大学を修了していない人が除去されやすい傾向にあることが示されており，本研究でも不良回答除去手法によって除去されやすい参加者の属性を確認する．

パート 2：セキュリティ関連の特徴を問う質問

パート 2 では，参加者のセキュリティ/ネットワークの知識を問う質問，および，セキュリティ行動の実施について問う質問を設けた．セキュリティ/ネットワークの知識を問う質問では，VPN や Cookie といった セキュリティ/ネットワーク用語 5 つに対して，参加者に自身の理解度を “No understanding” から “Full understanding” までの 5 段階で自己評価してもらった．参加者のセキュリティ行動の実施を問う質問では，セキュリティ行動評価指標である Security Behavior Intentions Scale（SeBIS） [7] を用いた．SeBIS は，デバイス管理方法やセキュリティアップデートの実施など，計 16 の質問項目から成る．参加者には各質問項目に対して，実施状況を “Never” から “Always” の 5 段階で自己評価してもらった．セキュリティ/ネットワークの知識，および，セキュリティ行動（SeBIS）に対する各参加者の回答を，各項目 1〜5 点として総合点を算出した．セキュリティ/ネットワークの知識の総合点は最低 5 点，最高 25 点で，セキュリティ行動（SeBIS）の総合点は最低 16 点，最高 80 点となる．

パート 3：フィッシングメール特定タスク

セキュリティ研究の中でも特にフィッシング研究においては，参加者の注意深さが結果に影響すると考えられ，不良回答除去手法によって注意力の高くない参加者を除去することが研究結果に大きく影響すると考えられる．そこで質問紙には，参加者にメールがフィッシングメールであるかを問う質問を設けた．参加者に提示するフィッシングおよび正規のメールは，既存のフィッシング研究 [5] のデータセットの中からランダムに抽出した．フィッシング特定タスクはロールプレイ形式を採用し，参加者にはメールの受信者のプロフィール情報を伝えた．参加者にはフィッシング 7 通，正規 7 通の計 14 通のメールのスクリーンショット（図 1）を提示した．メールの提示順は参加者によってランダマイズした．参加者にはフィッシングメールが含まれることは伝えたが，フィッシングメールの割合は伝えなかった．そして参加者に，各メールに対してフィッシングであると思うかどうかを “Yes”/“No” の二択で回答してもらった．

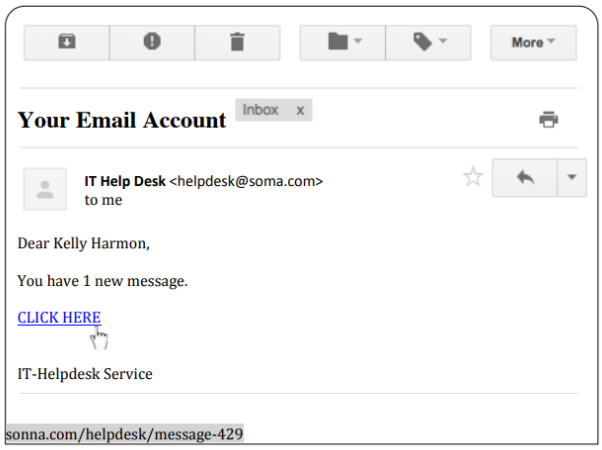

図 1 参加者に提示したフィッシングメールの例 [5]

3.2 不良回答除去手法

不良回答を行う参加者を検出・除去するにあたり，本研究では不良回答を行う参加者を以下のように 2 つのクラスに分類する．一つのクラスは「不誠実な参加者」で，調査への協力意欲が低い参加者を指す．具体的には，質問に対し事実を報告しようとしない参加者のこととする．このクラスにはボットによる自動回答を含む．もう一つのクラスは「不注意な参加者」で，調査項目に対する注意力が低かった参加者を指す．具体的には，質問文の精読を怠る参加者のこととする．

本研究では，不良回答除去手法の中でも学術研究でよく用いられる，5 種類の不良回答除去手法を用いた．以下にそれぞれの手法の概略を示す．

CAPTCHA

ボットによる自動回答（不誠実な回答）を除去するために，質問紙の冒頭に CAPTCHA を実装した．今回は，CAPTCHA を通過した回答のみを収集・分析した．

回答完了時間

Qualtrics の機能を用いて，参加者が全質問項目を回答するのに要した時間を取得した．本研究では，回答完了時間が質問量に対して不自然に短い場合，具体的には 3 分未満の場合に不良回答（不誠実な回答）と判定した．クラウドソーシングの場合は断続的にタスクに取り組むワーカー

も多い [9] ことから，回答完了時間が長い参加者の回答は除去しないこととした．

自由記述式質問

パート 2 の中に必須回答の自由記述式質問を 1 問設置した．今回は，自身の個人情報の管理ができている/できていないと思う理由を記入する質問とした．この質問に対し，明らかに回答になっていない回答を不良回答（不誠実な回答）とした．例えば，理由を聞いているのにも関わらず “Yes” “No” “None” “NA” などと記入している場合や，質問と関係のない文章を記入してしている場合などに不良回答とした．各回答が不良回答であるかどうかの判断は著者 2 名の体制で確認しながら実施した．

注意力テスト[*1]

パート 2 のセキュリティ行動を問うマトリクス形式の質問（SeBIS）の途中に注意力テストを設置した．これは，質問の代わりに，回答すべき選択肢を指示する文章（“Please select [OPTION] for this question.”）を提示し，指示に従っていない回答を不良回答（不注意な回答）として除去するものである．本研究で実装した注意力テストを以下に示す．

> Q. Please indicate how often you have done the following descriptions on the following scale: Never - Rarely - Sometimes - Often - Always.
>
> ⋮
>
> - I submit information to websites without first verifying that it will be sent securely (e.g., SSL, “https://”, a lock icon).
> - Please select “Never” for this question.
> - When browsing websites, I mouseover links to see where they go, before clicking them.
>
> ⋮

Instructional manipulation check（IMC）[*1]

Oppenheimer ら [11] によって提案された IMC は，上述の注意力テストと同様，回答すべき選択肢を指示することによって不良回答（不注意な回答）を除去する手法である．注意力テストは指示文章（“Please select ...”）のみが提示されるのに対し，一般に IMC では指示文章は見せかけの質問文の後に表示され，参加者は注意深く質問文全体を読まないと指示文章に気付くことができないように作られている．IMC の文字数が大きいほど不合格率が高くなる傾向にあり，より高い注意力をもつ参加者を厳選する結果となることが明らかになっている [2]．本研究で実装した IMC を以下に示す．

> Q. Have you ever attended an information-literacy lecture?
> Information literacy: determination of the extent of information needed, accessing the required information effectively and efficiently, evaluating information and its sources critically, incorporating selected information into one’s knowledge base, using information effectively to accomplish a specific purpose, and accessing and using information ethically and legally.
> This is a quality-check question, so please select the third option.

3.3 参加者募集

クラウドソーシングサービスの Amazon Mechanical Turk（MTurk） [1] を利用して参加者募集を実施した．アメリカ在住の 18 歳以上のワーカー（MTurk 登録者）を 300 人募集した．さらに今回は，多くの学術研究と同様に，過去のタスク承認率が 95%以上のワーカーのみに募集を限定した．報酬額は，パイロットテストにおける参加者の平均回答完了時間が 11.8 分であったことを踏まえ，時給換算でアメリカの最低賃金を優に超える 2.4 米ドルとした．

研究倫理の一環として，募集時にはインフォームドコンセントを実施した．具体的には，調査内容・データの取り扱い方法・想定所要時間・報酬額などをワーカーに伝えた上で，調査参加に同意した参加者のみにアンケートサイト（Qualtrics）に進んでもらった．なお，分析時に不良回答と分類されるかどうかに関わらず，全質問項目への回答が確認できた全ての参加者に対して報酬を支払った．

4. 調査結果

本章では, 本調査で得られた結果について述べる. 具体的には，CAPTCHA を通過した参加者を他の不良回答除去手法によって誠実さ・注意力の程度を 4 グループに分類し，各グループの回答結果を比較した．

4.1 不良回答除去手法による参加者の分類

CAPTCHA を通過した参加者 300 名を，回答完了時間・自由記述式質問・注意力テスト・IMC の回答内容によって 4 グループに分類した．分類方法および分類結果を表 1 に示す．なお，注意力テストには通過しなかったが IMC を通過した 1 人の参加者に関しては，今回は分析対象外とした．

表 1 より，MTurk における過去のタスク承認率が 95%以上のワーカーに参加を制限し，かつ，CAPTCHA を通過した参加者の回答のみを収集したのにも関わらず，40.3%もの参加者が不誠実な参加者に分類されたことが見て取れる．不誠実と分類された参加者のうち，82.8%の参加者は自由記述式質問への回答のみが不合格，6.6%の参加者者は回答完了時間のみが不合格，10.7%の参加者は回答完了時

[*1] 注意力テストおよび IMC は，文献により指す内容が異なる場合があるが，本稿では記載の通りに定義した．

表 1 参加者の分類方法および分類結果

参加者の分類	完了時間	自由記述	注意力テスト	IMC	割合
グループ 1．不誠実な参加者	一方または両方が×		–	–	40.3%
グループ 2．誠実だが注意力が低い参加者	○	○	×	×	0.0%
グループ 3．誠実だが注意力が中程度の参加者	○	○	○	×	11.6%
グループ 4．誠実で注意力が高い参加者	○	○	○	○	47.9%

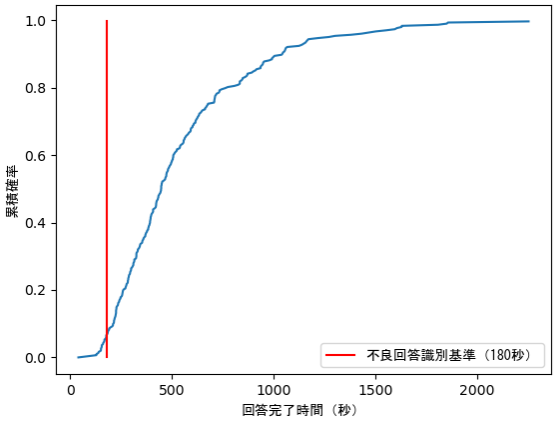

図 2 回答時間の累積分布

間と自由記述式質問の両方が不合格であった．不良回答と分類された自由記述式回答の中には，インターネット上の何らかの文章をコピー＆ペーストしただけと思われる回答が複数見つかった．グループ 2（誠実だが注意力が低い）に分類された参加者は 0 人であった．本調査においては，不誠実な参加者の除去を想定して自由記述式質問を設置したが，自由記述式質問は注意力テストと同等もしくはそれ以上に，不注意な参加者の除去にも作用する可能性が示唆された．誠実であると判断された参加者の中では，IMC を通過し，誠実で注意力が高い（グループ 4）と分類される参加者が多かった．

なお，アンケート全体の平均回答完了時間は，グループ 1 が 7.4 分，グループ 3 が 9.6 分，グループ 4 が 10.4 分であり，誠実で注意力が高い参加者ほど時間をかけて回答していた．参加者の回答完了時間の累積分布を図 1 に示す．

次に，各グループに分類された参加者のデモグラフィーを表 2 に示す．先行研究 [8] で示されたのと同様に，男性・大学を修了していない参加者のほうが IMC により除去されやすい傾向にあった．先行研究 [8] では若い参加者のほうが IMC により除去されやすい傾向にあったが，本研究ではそのような傾向は確認できなかった．また，IT 関連の職歴がなく，スマートフォン利用時間が短い参加者のほうが IMC により除去されやすいことも明らかになった．このように，研究者が適用する不良回答除去手法によって参加者のデモグラフィーに偏りが生じるため，研究者は不良回答除去手法の適用について慎重に検討する必要がある．

4.2 セキュリティ知識・行動

参加者のセキュリティ/ネットワークの知識，およびセキュリティ行動（SeBIS）の得点を表 3 に示す．誠実な参加者であったグループ 3 とグループ 4 では，セキュリティ知識に有意差は見られなかった（t 検定，p =0.26）が，セキュリティ行動では注意力が中程度の参加者の方が高程度の参加者よりも有意に総合点が高かった（t 検定，p <0.05）．ただし，セキュリティ知識・行動の総合点は，客観的評価指標ではなく自己評価に基づくものであるため，検証にはさらなる調査が必要である．参加者のセキュリティ知識とフィッシングメール特定パフォーマンスの関係については，4.3 節で考察を行う．

4.3 フィッシングメール特定タスクの正答率

参加者のフィッシングメール特定タスクの正答率を比較した結果を表 4 に示す．ここで，「総合正答率」は計 14 通のメールのうち正規/フィッシングを正しく特定できた率，「正規正答率」は計 7 通の正規メールのうち正規であるとを正しく特定できた率，「フィッシング正答率」は計 7 通のフィッシングメールのうちフィッシングであると正しく特定できた率を指す．「タスク完了時間」はフィッシングメール特定タスクの完了に要した時間を指す．

グループ 1 の不誠実な参加者の結果は，総合正答率，正規正答率，フィッシング正答率，タスク完了時間のすべての項目において，グループ 3 やグループ 4 の誠実な参加者の回答から乖離していることがわかる．タスク完了時間の短さ，正規正答率の低さ，フィッシング正答率の高さからして，グループ 1 には，提示されたメールを吟味することなく “Yes”（フィッシングメールであると思う）を回答する参加者が多いことが推測される．この結果は，グループ 1 の不誠実な参加者の回答の多くはセキュリティのユーザスタディとしては除外すべきノイズであることを示唆している．すなわち，**回答完了時間および自由記述式質問による不良回答除去には一定の効果がある**ことが示唆される．

次に，グループ 3 の注意力が中程度の参加者とグループ 4 の注意力が高い参加者の結果を比較する．総合正答率やタスク完了時間に大きな差は見られなかった一方で，正規正答率とフィッシング正答率にはグループ間で有意な差が見られた（t 検定，正規正答率：p <0.01，フィッシング正答率：p <0.05）．セキュリティ行動の得点はグループ 3 の方が良いにも関わらず，注意力が高い参加者のほうが注意力が中程度の参加者よりも正規正答率が低いこと，および，

表 2 各グループに分類された参加者のデモグラフィー

参加者の分類	男性の割合	34 歳以下の割合	大学修了者の割合	IT 職歴有の割合	スマホ利用 3 時間以上の割合
グループ 1（不誠実）	52.5%	62.3%	95.1%	95.1%	66.4%
グループ 2（注意力低）	–	–	–	–	–
グループ 3（注意力中）	74.3%	60.0%	60.0%	34.3%	28.6%
グループ 4（注意力高）	64.1%	60.0%	77.2%	60.0%	54.5%
参加者全体	60.4%	61.1%	82.5%	71.3%	56.4%

表 3 セキュリティ知識・行動の回答結果

参加者の分類	総合点（平均）	
	セキュリティ知識	セキュリティ行動
グループ 1（不誠実）	17.6	51.9
グループ 2（注意力低）	–	–
グループ 3（注意力中）	18.6	61.0
グループ 4（注意力高）	17.7	57.5
参加者全体	17.8	55.6

注意力が高い参加者のほうが注意力が中程度の参加者よりもフィッシング正答率が高いことから，アンケート調査中の注意力が高い参加者のほうがメールをフィッシングだと判断する傾向にあることがわかった．以上から，参加者の注意深さの程度と参加者のフィッシングメール特定傾向には関連があること，すなわち**フィッシングに関するユーザスタディにおいて，安直な IMC の適用は研究結果に偏りを生じさせる原因となり得る**ことが明らかになった．メールを正規と判断する傾向がより強い参加者の方がフィッシングの被害に遭うリスクが高いため，フィッシング研究においてはグループ 3 の参加者の認識や行動を理解することが重要であると考えられる．

最後に，参加者のセキュリティ行動とフィッシングメール特定正答率の関係について考察する．表 4 より，グループ 3（注意力が中程度）の参加者の方がグループ 4（注意力が高い）の参加者よりもフィッシングの被害に遭うリスクが高い傾向がみてとれた．一方，表 3 では，グループ 3 とグループ 4 の参加者を比較すると，両者は同程度のセキュリティ知識をもつが，グループ 3 の参加者がより多くのセキュリティ行動を実践している傾向にあった．すなわち，フィッシングメール特定の正答率は，参加者のセキュリティ知識・行動よりも，メール対応時点での参加者の注意力に大きく影響を受ける可能性も示唆された．この場合，フィッシングメール対策として，ユーザのセキュリティ知識・行動を向上させるようなアプローチに比べ，メール対応時のユーザの注意力を一時的に高めるフィッシングメール警告を提示するようなアプローチのほうが即効性がある可能性がある．前述したように，セキュリティ知識・行動の総合点が自己評価に基づくことに由来する可能性があるため，確固たる結論を導くには慎重に追加分析を実施することが必要である．

5. 議論

5.1 不良回答除去手法の有効性と課題

以下では，本研究で得られた結果をもとに，オンラインユーザ調査における不良回答除去手法の有効性と 2 つの課題を議論する．

【有効性】

MTurk を用いた実験の結果，過去のタスク承認率 95% 以上のワーカーに限定して参加者を募集したにも関わらず，40%が不誠実な参加者であると判定された．クラウドソーシングサービスにおけるワーカーの評価（過去のタスク承認率・承認数）は，そのワーカーの回答の質を必ずしも反映するものではないことが示唆された．4.3 節で述べた通り，回答完了時間および自由記述式質問によって不誠実であると分類された参加者の回答内容は実際に質が低く，オンラインユーザ調査における不良回答除去手法が有効であることが示された．本研究では特に自由記述式質問による不良回答除去の有効性が確認されたが，自由記述式質問への回答の分析は研究者による手作業で実施するため，研究者の経験やスキルに依存する点に注意を要する．また，自由記述式質問への回答の分析は研究者の労力を要するため，大規模調査に対する適用が困難となる問題もある．

【課題 1】

一つの課題は，不良回答除去手法の適用により，有効となる参加者のデモグラフィーに偏りが生じる懸念である．本研究では，IMC を適用することによって男性および大学を修了していない人が除去されやすいことが確認された．この傾向は先行研究 [8] の結果と一致している．こうした IMC の適用がもたらす性別・学歴の偏りは，幅広くオンライン調査に基づくセキュリティ研究の結果に影響を及ぼすと考えられる．アンケート作成サービスの Qualtrics や学術研究向けクラウドソーシングサービスの Prolific Academic [13] では，参加者のデモグラフィーの偏りへの懸念，およびワーカー保護の観点から，IMC の適用を非推奨としている [14], [16][*2]．

【課題 2】

もう一つの課題は，IMC などの注意力の高くない参加者を除去する不良回答除去手法を適用することで，セキュリティ研究において本来対象とすべき参加者が除去される懸

[*2] 厳密には，Prolific では，研究者が IMC の結果をもとにワーカーを否認しワーカーに報酬を支払わないことを禁止としている．

表 4　フィッシングメール特定タスクの正答率

不良回答除去基準	総合正答率（平均）	正規正答率（平均）	フィッシング正答率（平均）	タスク完了時間（平均）（分）
グループ 1（不誠実）	50.9%	29.6%	72.1%	2.14
グループ 2（注意力低）	–	–	–	–
グループ 3（注意力中）	64.3%	76.3%	52.2%	3.84
グループ 4（注意力高）	61.5%	59.4%	63.6%	3.99
参加者全体	57.5%	49.3%	65.7%	3.22

念である．人間を対象としたセキュリティ研究では，ユーザの注意深さはユーザの認識や行動を理解する上で重要な要因となりうる．特に，参加者の注意深さが結果に大きく影響すると考えられるフィッシング研究においては，注意力の程度による参加者の除去は不適切である可能性がある．我々は，IMC を適用することによって，メールを正規と判断しやすい参加者，つまり，フィッシング被害に遭うリスクの高い参加者が除去されやすい傾向があることを明らかにした．しかしながら，一般に人間の注意力は時間とともに変化するものであり，単一のユーザ調査における注意力テストや IMC の結果はその参加者の恒常的な性質としての注意深さの程度を必ずしも表すものではないため，調査結果の解釈には注意が必要である [2], [3].

本研究の結果は，オンラインユーザ調査を伴う研究において研究者がどのように不良回答除去を実施するかの選択が研究結果に大きく影響を与える危険性を示唆している．研究者は研究目的に応じて慎重に不良回答除去の検討を行う必要がある．

5.2　制約事項と今後の研究課題

本研究は不良回答除去手法の適用によりセキュリティ研究の結果に偏りが生じ得ることを実証した初の研究である一方で，下記に示す 3 つの制約がある．以下，それぞれの制約と，今後の課題を示す．

【制約 1】

本調査では 5 種類の不良回答除去手法を 1 回のアンケート調査の中に設置したため，参加者は注意力テストと IMC の 2 種類の不良回答除去を直接目にした．これにより，参加者の注意力は通常のタスク実行時より上がってしまった可能性がある．また，実装した CAPTCHA・注意力テスト・IMC の設置位置は固定しており，設置位置による影響を測定できていない．今後，不良回答除去手法の設置個数や設置位置を変え，本調査と同様の結果が得られるかを検証する．

【制約 2】

本調査で実装した不良回答除去手法は 5 種類のみであるが，他にも様々な不良回答除去手法が実際の学術研究において適用されている．今後は，回答分布，クリック数，IP アドレス，トラップ質問，一貫性チェック質問などの不良回答除去手法についても検証を行い，各種不良回答除去手法の実装による影響をまとめたい．また，現在のクラウドソーシングサービスにおいては，検出回避を狙った高度なボットの存在も考えられるため，より高精度に不良回答を除去する方法を明らかにする必要がある．

【制約 3】

本調査では MTurk のみで参加者募集を実施した．学術研究においては MTurk と Prolific Academic がよく用いられるが，MTurk と Prolific Academic では，登録ワーカーの特徴が異なることが明らかになっている [12]．具体的には，ワーカーの居住国やワーカー 1 人当たりのタスクの実施数の程度が異なる．今後は Prolific Academic でも同様の調査を行い，クラウドソーシングサービスによって不良回答除去手法の影響が異なるかどうかを明らかにする必要がある．

6.　結論

本研究では，オンラインユーザ調査における不良回答除去手法の実装がセキュリティ研究，特にフィッシング研究の結果に及ぼす影響について調査を行った．フィッシングメール特定に関するオンライン調査に不良回答除去手法を導入した結果，不誠実な参加者が全体の 40%を占めること，および，回答完了時間および自由記述式質問による不良回答除去は有効性が認められることを明らかにした．また，IMC による不良回答除去により参加者のデモグラフィーに偏りが生じる問題，およびユーザの注意深さが結果に大きな影響を与える研究において，IMC の適用は研究結果の偏りにつながり得る問題があることを示した．本研究の結果は，オンラインユーザ調査を伴う研究において研究者がどのように不良回答除去を実施するかの選択が研究結果に大きく影響を与える危険性を示唆している．今後は，今回調査した 5 種類以外の不良回答除去手法についても有効性や及ぼす影響の調査を実施する．

参考文献

[1] Amazon: Amazon Mechanical Turk, `https://www.mturk.com/`. (参照 2021-01-22).
[2] Anduiza, E. and Galais, C.: Answering without reading: IMCs and strong satisficing in online surveys, *International Journal of Public Opinion Research*, Vol. 29, No. 3, pp. 497–519 (2017).
[3] Berinsky, A. J., Margolis, M. F. and Sances, M. W.: Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys, *American Journal of Political Science*, Vol. 58, No. 3,

pp. 739–753 (2014).
[4] Buchanan, E. M. and Scofield, J. E.: Methods to detect low quality data and its implication for psychological research, *Behavior Research Methods*, Vol. 50, No. 6, pp. 2586–2596 (2018).
[5] Canfield, C. I., Baruch, F. and Alex, D.: Quantifying phishing susceptibility for detection and behavior decisions, *Human factors*, Vol. 58, No. 8, pp. 1158–1172 (2016).
[6] Dickinson, D. L. and McEvoy, D. M.: Further from the Truth: The Impact of In-Person, Online, and mTurk on Dishonest Behavior, *IZA Discussion Paper*, No. 13686 (2020).
[7] Egelman, S. and Peer, E.: Scaling the security wall: Developing a security behavior intentions scale (sebis), *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (2015).
[8] Kapelner, A. and Chandler, D.: Preventing satisficing in online surveys: A "Kapcha" to Ensure Higher Quality Data, *Proceedings of the CrowdConf'10* (2010).
[9] Lascau, L., Gould, S., Cox, A., Karmannaya, E. and Brumby, D.: Monotasking or Multitasking: Designing for Crowdworkers' Preferences, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).
[10] Moss, A. J. and Litman, L.: After the bot scare: Understanding what' s been happening with data collection on MTurk and how to stop it, *Retrieved February*, Vol. 4, p. 2019 (2018).
[11] 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).
[12] Oppenlaender, J., Milland, K., Visuri, A., Ipeirotis, P. and Hosio, S.: Creativity on Paid Crowdsourcing Platforms, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).
[13] Prolific: Prolific Academic, `https://www.prolific.co/`. (参照 2021-01-22).
[14] ProlificTeam: Using attention checks as a measure of data quality, `https://researcher-help.prolific.co/hc/en-gb/articles/360009223553-Using-attention-checks-as-a-measure-of-data-quality`. (参照 2021-01-22).
[15] Qualtrics: Qualtrics, `https://www.qualtrics.com/`. (参照 2021-01-22).
[16] Vannette, D.: Using Attention Checks in Your Surveys May Harm Data Quality, `https://www.qualtrics.com/blog/using-attention-checks-in-your-surveys-may-harm-data-quality/` (2017). (参照 2021-01-22).
[17] Yarrish, C., Groshon, L., Mitchell, J., Appelbaum, A., Klock, S., Winternitz, T. and Friedman-Wheeler, D. G.: Finding the signal in the noise: Minimizing responses from bots and inattentive humans in online research, *The Behavior Therapist*, Vol. 42, No. 7, pp. 235–242 (2019).