Careless Participants Are Essential for Our Phishing Study: Understanding the Impact of Screening Methods

Tenga Matsuura
Waseda University
tenga1012@nsl.cs.waseda.ac.jp

Ayako A. Hasegawa
NTT

Mitsuaki Akiyama
NTT

Tatsuya Mori
Waseda University / NICT / RIKEN AIP

ABSTRACT

Online surveys using crowdsourcing services have been widely adopted in academic research projects aimed at understanding human perception and behavior. Because there is a concern that online surveys may include dishonest or careless responses by crowdworkers who perform a large number of tasks, or responses by bots, several screening methods have been proposed to discard such low-quality responses. However, in security research, especially in phishing research where the attention of participants is considered to influence the results, the elimination of careless 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 screening methods bias the results of security surveys?*" Using Amazon Mechanical Turk and Prolific Academic, two popular crowdsourcing platforms used in online surveys, we conducted online user studies ($N = 600$) on security knowledge, security behavior, and phishing email detection performance to elucidate the influence of screening methods on the results. The obtained results indicate that the adoption of the instructional manipulation check (IMC) screening method triggers bias in the demographics of the participants, as well as differences in the results of phishing email detection performance. In addition, the degree of these differences depends on the crowdsourcing platform. We also demonstrated that it is non-trivial to determine the correlation between screening methods and factors that can influence the results of a survey on security behavior. These findings suggest that caution should be exercised when applying screening methods such as attention checks and IMC in studies where the extent of user attention could have a significant impact on the results.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**.

EuroUSEC '21, October 11–12, 2021, Karlsruhe, Germany

ACM ISBN 978-1-4503-8423-0/21/10…$15.00
https://doi.org/10.1145/3481357.3481515

KEYWORDS

Crowdsourcing, Phishing, Attention Check, Instruction Manipulation Check

ACM Reference Format:
Tenga Matsuura, Ayako A. Hasegawa, Mitsuaki Akiyama, and Tatsuya Mori. 2021. Careless Participants Are Essential for Our Phishing Study: Understanding the Impact of Screening Methods. In *European Symposium on Usable Security 2021 (EuroUSEC '21), October 11–12, 2021, Karlsruhe, Germany*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3481357.3481515

1 INTRODUCTION

Academic studies aimed at understanding human cognition and behavior often adopt the user survey approach. Conventionally, telephone, letters, or in-person interviews have been employed in user surveys. Recently, the use of online surveys has been on the rise, and with the impact of COVID-19, its adoption is expected to increase even further. Although online user surveys have various advantages such as the ability to recruit a wide range of participants, inherent factors adversely affect the quality of the data. In general, crowdsourcing workers can get paid more if they complete a large number of tasks in a short period of time; such a model will incentivize some workers to answer questions carelessly without examining them and/or adopting assistive tools that support fast/automated response input. Therefore, in some cases, the results of online surveys contain careless or dishonest responses.

An approach to eliminating such careless or dishonest responses is to adopt screening methods. Among the various types of screening methods, instructional manipulation check (IMC), which was proposed by Oppenheimer et al. [22] is the most powerful and widely adopted method. IMC has a deceptive aspect because it is designed in such a way that questions cannot be answered without reading carefully, and the content cannot be fully understood at first glance. By applying IMC, the level of inattention can be quantified, and low-quality responses can be discarded from subsequent analyses. Hauser et al. [16] conducted an experiment on Amazon mechanical Turk (MTurk) workers using IMC, and determined that MTurk workers paid more attention to instructions than college students, thus suggesting that using MTurker pool samples may yield better results than adopting conventional subject pool samples in social science research. IMC is employed in wide range of research fields, and usable security and privacy research is no exception. In several papers presented at SOUPS, which is a prominent conference in the research field, researchers have employed IMC in their questionnaires to ensure data quality [4, 9].

The known disadvantage of employing IMC is that it could result in demographic bias, as participants with certain characteristics are removed [5, 18]. Berinsky et al. [5] demonstrated that when IMC was implemented, there was a bias toward older people and women exhibited higher success rates. In the security research context, if we were to conduct a user study on security behavior against phishing websites, we would expect the attentive level of the participants to be a factor that would significantly influence the results. Therefore, in such a study, removing participants with low attentiveness may incur the risk of drawing conclusions that are partly erroneous.

Based on these backgrounds, this study addresses the following research question (RQ).

RQ: *Does the adoption of survey response screening methods introduce bias into the results of a user survey on security behavior?*

Here, we will focus on CAPTCHA, response completion time, open-ended responses, attention checks, and IMC as screening methods for survey responses.

To address the RQ, this study adopts the screening method described above to classify participants into groups, according to their level of honesty and attention. Subsequently, we compare the security knowledge, security behavior, and phishing email detection performance of participants in each of the classified groups.

To ensure the generality of our findings, it is essential to obtain results on different crowdsourcing platforms. In this study, we conduct a user experiment using two leading crowdsourcing services adopted in online survey research: MTurk and Prolific Academic [25]. We employed 300 participants each for a total of 600 participants. As will be demonstrated later, we report that workers on each platform exhibit significantly different characteristics.

We also demonstrate that the adoption of screening methods based on the attention level of the participants triggers intrinsic demographic bias. Furthermore, we report that the participants with medium and high attention levels clearly differed in their tendency to make judgments on phishing e-mails, thus implying that applying an attention-based screening scheme will hide these difference, which could be a crucial factor that determines user behavior. Based on the results obtained in this study, we discuss the appropriate implementation of screening methods in user studies for security research.

The main contributions of this paper are as follows

- This is the first study to demonstrate that applying screening methods to test participants' attentiveness can bias the demography and results of user surveys on security behavior.
- We demonstrated the need for researchers to be extra cautious when applying screening methods such as attention check and IMC in studies where the extent of user attention significantly impacts the results.

The structure of this paper is as follows. First, we summarize the background of our research in Section 2. Then, we describe our experimental methodologies in Section 3, and the results of the experiments in Section 4. In Section 5, we first discuss the effectiveness and challenges of the screening methods based on the results obtained, and then discuss the limitations of this study and future research directions. Finally, Section 6 summarizes the findings of this study.

2 BACKGROUND AND RELATED WORK

In this section, we review studies that discuss the impact of low-quality responses in online surveys using crowdsourcing platforms, as well as the studies that discuss screening methods for eliminating such low-quality responses.

Amazon Mechanical Turk (MTurk) [2] and Prolific Academic [25] have become widely used in academic research, for example, in online surveys. Compared to conventional surveys, online surveys have the advantage of being able to recruit a large number of diverse participants in a short period of time, which is highly convenient for researchers. Redmiles et al. studied the generalizability of MTurk surveys on both privacy and security behavior, and determined that MTurk respondents tend to mirror the weighted probabilistic sample representative of the entire U.S. population [27]. However, an inherent problem with online surveys is that they include dishonest or careless responses by workers who sacrifice the quality of the task and prioritize completing a large number of tasks in a short period of time. In particular, it has been reported that in MTurk, there has been an abrupt increase in low-quality responses since 2018 [20]. Dickinson et al. observed that the factors contributing to low-quality responses in online surveys include the monetary compensation model for participants, as well as the anonymous response format and lack of online research monitoring [10].

To eliminate low-quality responses in online surveys, various types of screening methods have been proposed and tested for their effectiveness, especially in the field of social psychology [6, 7, 12, 22, 29, 31]. For example, Yarrish et al. [31] investigated the disparity between human and bot-based responses in online surveys, using various screening methods. The results obtained indicated that although CAPTCHAs can completely eliminate automated bot responses, they are faced with challenges in eliminating the responses of human-in-the-loop bots. They also observed that attention check is less effective for experienced workers who only answer such tests correctly. In addition, they observed that although employing open-ended questions is effective in extracting legitimate responses, they come at a higher cost to the researcher. Buchanan et al. [7] investigated the effect of combining multiple tests, such as click count, page timing, number of scale options, data distribution, and manipulation check, and verified their effectiveness. Based on their experiments, they recommended that a response should be discarded if it is flagged in two of the five tests.

While several social psychology researchers have developed screening methods, such as CAPTCHA, open-ended responses, attention checks, and IMC, security researchers have also adopted the methods [4, 9]. However, as mentioned in Section 1, there is a risk that the implementation of IMC may result in the removal of participants with certain characteristics, which triggers demographic bias [5, 18]. Kapelner and Chandler conducted an online survey on MTurk and demonstrated that females, older people, and college graduates were more likely to pass IMC [18]. Berinsky et al. conducted an online survey via a web survey company and obtained similar results to the research of Kapelner and Chandler, as they demonstrated that females, older people, and Caucasians

were more likely to pass IMC [5]. Based on the results of these studies, Qualtrics [26], a company that provides comprehensive tools for creating online surveys and analyzing collected data, warns their customers about the risk to survey results owing to attention checks [30].

We assumed that the implementation of a screening method that removes participants who are not highly attentive may lead to the removal of participants who should be studied in security research, especially in research aimed at understanding user cognition and behavior toward phishing attacks, where the attentiveness of the participants may affect the results. In this study, we investigated whether applying existing screening methods could bias the results of a security research questionnaire.

Finally, we present a comparison between two widely used online research platforms: MTurk and Prolific Academic [23, 24]. MTurk workers are mostly from the US, followed by India, while Prolific workers are primarily from the UK, followed by the US and other countries. The average age of MTurk workers is 33.3 (σ = 8.9), whereas that of Prolific is 37.1 (σ = 11.6). Regarding weekly work hours, prolific workers are a much more casual workforce than MTurk workers. In addition, the time taken to recruit participants was shorter for MTurk. Adams et al. compared data quality from MTurk and Prolific and demonstrated that Prolific workers' completion rates, diversity, attention, naivety, reproducibility, and honesty are better than MTurk workers [1].

3 DESCRIPTION OF THE EXPERIMENTS

In this study, we conducted an experiment using a crowdsourcing service to verify whether the implementation of a screening method in an online survey can bias the results of a security behavior study. In the following, we describe the questionnaire used in the experiment, the details of the screening method adopted in the survey, and the procedure for recruiting participants.

3.1 Questionnaire

The questionnaire we developed is divided into three parts as follows. **Part 1**: Questions about the demography of participants, **Part 2**: Questions on the security knowledge, behavior, etc., and **Part 3**: Phishing email detection task. In total, there are 53 questions asked, and multiple screening methods are inserted (See appendix for the full version of the questionnaire). The structure of each part is as follows. Part 1 consists of 17 questions, including one open-ended question and one IMC, Part 2 consists of 22 questions, including one attention check, and Part 3 consists of 14 e-mail classification tasks, including 7 legitimate emails and 7 phishing emails. The questionnaire is written in English, and the experimental participants adopt Qualtrics [26] as a platform to submit their responses to the questions.

Part 1: Demographic questions

In Part 1, we designed demographic questions that asked participants about their age, gender, last education, IT work experience, device used, and time spent using the device. A previous study [18] conducted by Kapelner et al. in 2010 demonstrated that the application of IMC tends to discard the responses of younger people, males, and people without a college degree. We conducted a similar study, using MTurk and Prolific to determine whether such results are reproducible.

Part 2: Questions about security knowledge and behavior

In Part 2, we asked the participants questions about their technical knowledge of security and security behavior. For the questions on the knowledge of security technology, we asked the participants to evaluate their own understanding of five technical terms, i.e., IP address, malware, SSL/TLS, VPN, and cookies, on a five-point scale from "No understanding," "Little understanding," "Some understanding," "Good understanding," and "Full understanding." These items are based in part on a previous study [14].

The security behavior intentions scale (SeBIS) [13], a security behavior evaluation index, was adopted for the questions on the security behavior of participants. SeBIS consists of 16 questions, such as device management methods and security update practices. Participants were asked to self-evaluate their practices on a five-point scale from "Never" to "Always" for each question. The overall score was calculated from each participant's responses to the security technology knowledge and security behavior (SeBIS), with each item scored from 1 to 5. The total scores for security knowledge were a minimum and maximum of 5 and 25 points, while those of security behavior (SeBIS) were 16 and 80 points, respectively.

Part 3: Phishing Email Detection Task

In Part 3, participants were asked to answer whether the presented email was a phishing email. This task is intended to measure a user's ability to detect phishing. By correlating and analyzing the results of this measurement with the results of applying the screening method, we can evaluate the relationship between the attention level and phishing detection ability. The phishing and legitimate emails presented to the participants were randomly selected from the dataset of an existing phishing study [8]. The phishing email detection task adopted a role-playing format, and participants were provided with the profile information of the email recipients. The roleplay task enables researchers to study the behaviors and perceptions of participants to phishing attacks without conducting an actual simulated phishing attack [28]. We note that there is room to consider the ecological validity of the roleplay tasks in our online survey, for example, classifying emails using their screenshots. However, exploring the design space for more ecologically valid phishing tasks is beyond the scope of our study.

Participants were presented with screenshots of 14 emails: 7 phishing emails and 7 legitimate emails. The number of emails used in the task exceeds that of prior research based on phishing roleplay task, e.g., Downs et al. [11] used eight emails. Figures 1 and 2 show screenshot examples of a legitimate and phishing email, respectively. Summary of emails used in the experiment is shown in Table 5. The order of the emails was randomized by the participants. We informed the participants that phishing mails were included; however they were not informed about the percentage of phishing emails. Subsequently, we asked the participants to answer "Yes"/"No" to whether they thought each email was a phishing email.

3.2 Screening Methods

In this study, we evaluated five screening methods that are commonly adopted in survey research. A brief description of each method is provided below. We note that the terms "attention check"

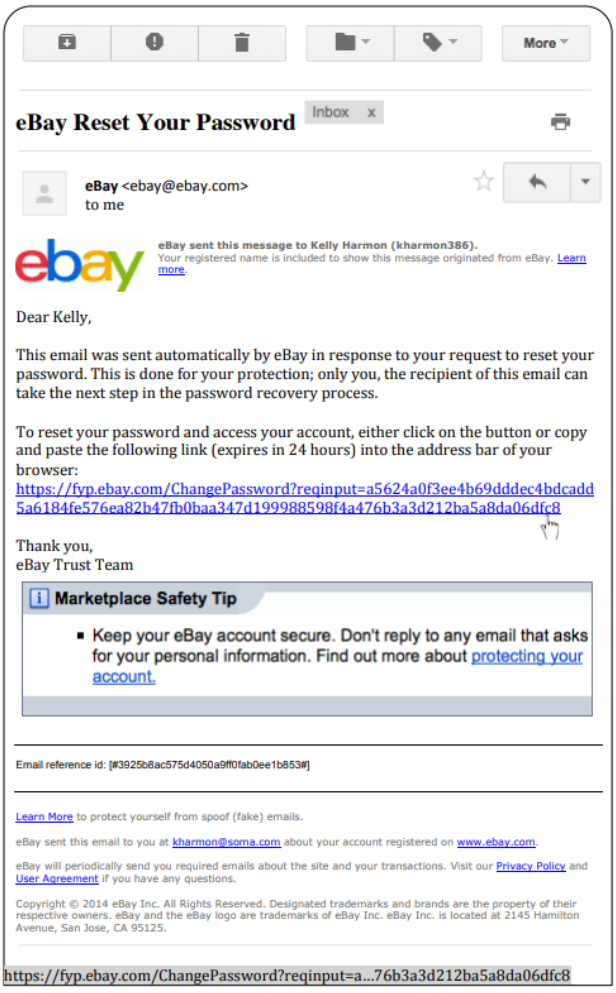

Figure 1: An example of a legitimate email presented to the participants [8].

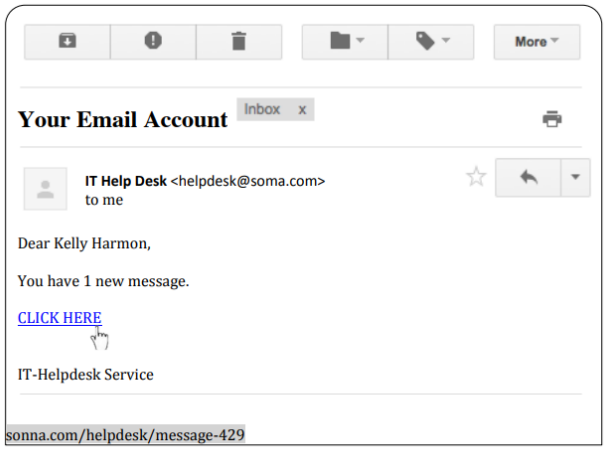

Figure 2: An example of a phishing email presented to the participants [8].

and "IMC" may have different meanings in different literature. In this study, they are defined as follows:

CAPTCHA.

A CAPTCHA is a type of challenge/response test used to verify that the responder is not a machine. A typical CAPTCHA system aims to distinguish between humans and machines based on whether or not the respondent can read the letters and numbers in the image. In this study, we placed a CAPTCHA at the beginning of the questionnaire to discard automated responses by bots. We solely collected and analyzed responses that passed the CAPTCHA test.

Response Completion Time.

This study measured the time taken by participants to answer all questionnaire items. In this study, participants were considered to be "dishonest respondents" if their response completion time was unnaturally short in relation to the number of questions asked, specifically, if the time taken was less than 3 minutes. The reason for setting the time as less than 3 minutes is that the pilot test result showed difficulty in understanding and completing our online survey less than 3 minutes. The method for classifying participants is presented in the next section. In crowdsourcing, there are many workers who work on tasks intermittently [19]; hence, we decided not to discard responses from participants with long response completion times.

Open-ended questions.

An open-ended question was included in Part 2. The question requested the respondents to explain why they thought they were or were not able to manage the protection of their own personal information. If the answer to this question is obviously not valid, it is considered a dishonest answer and discarded. For example, if they respond with "Yes," "No," "None," "NA," etc., even though the question was asked for a reason, or if they wrote sentences that had nothing to do with the question, we judged the answer as not valid. The validity of each response was ascertained by two authors who mutually verified the results of their judgments.

Two coders, who were the authors, independently reviewed open-ended responses. Each coder spent approximately 4 h classifying the responses into honest or dishonest for approximately 600 open-ended responses. The codes were compared, and inter-rater reliability was calculated using Cohen's kappa coefficient. The inter-rater reliability was 0.88, which is considered to be in high agreement.

Attention check.

In Part 2, an attention check item was placed in a matrix of the Likert items for the SeBIS questions, with the aim of measuring security behavior. This check does not pose a question, instead it presents a sentence that instructs the choices to be answered, such as "Please select [OPTION] for this question." If the answer does not follow the instructions, it is judged as an answer with low attention. The attention check implemented in this study is illustrated in Figure 3.

Table 1: Definitions of participants groups.

Groups	Descriptions	Completion time	Open-ended	Attention check	IMC
G1	Dishonest participants	Failure in one or both		–	–
G2	Honest participants with low attention	Pass	Pass	Fail	Fail
G3	Honest participants with moderate attention	Pass	Pass	Pass	Fail
G4	Honest and attentive participants	Pass	Pass	Pass	Pass

Q. Please indicate how often you have performed the following descriptions on the following scale: Never- Rarely- Sometimes- Often- Always.

⋮

- I submit information to websites without first verifying that they will be sent securely (e.g., SSL, "https://," and a lock icon).
- Please select "Never" for this question.
- When browsing websites, I mouseover links to see where they go before clicking them.

⋮

Figure 3: An example of attention check.

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.

Figure 4: An example of IMC.

Instructional manipulation check (IMC).

IMC, which was proposed by Oppenheimer et al. [22], is a method for detecting inattentive responses by indicating the choices to be answered, which is similar to the attention check described above. In contrast to the attention check, where only the instructional text is presented, in IMC, the instructional text is generally presented after the dummy question text, such that the participant has to read the entire question text more carefully to notice the instructional text. It has been reported that the higher the number of letters in the IMC, the higher the probability that the participant will not pass the screening [3]. In other words, by applying an IMC, it is possible to carefully identify participants with high attention levels. The IMC implemented in this study is presented in Figure 4.

3.3 Classification of participants

To efficiently detect participants with dishonest or careless responses, the participants were categorized into four groups, as presented in Table 1. **Group 1** is "dishonest participants," which refers to participants that are not willing to cooperate with the true purpose of the survey, specifically, those who do not want to report facts in their response to questions. This group includes responses obtained with tools and/or bots. In this study, participants were classified as **Group 1** if their completion time was less than 3 min, or if their responses to open-ended questions were judged to be invalid. **Group 2** is "honest participants with low level of attention." Participants were classified as **Group 2** if they passed the completion time and open-ended screening, but did not pass either the attention test or the IMC. **Group 3** is "honest participants with a moderate level of attention." If a participant passes the completion time, open-ended, and attention test screenings, but does not pass the IMC, we classify them as **Group 3**. Finally, **Group 4** is "honest and attentive participants." If a participant passes all the screenings, we classify them as **Group 4**.

3.4 Recruiting participants

We employed two crowdsourcing platforms, MTurk [2] and Prolific Academic [25], to recruit participants for our experiment. In MTurk, we recruited 300 workers aged over 18 years living in the US. As in most academic studies, we restricted the recruitment to only those workers with a past task approval rate exceeding 95% Based on the observation that the average completion time of participants in the pilot test was 11.8 min, we set the reward of our task at 2.4 USD, which is well above the minimum wage in the US [21].

For the recruitment of participants in Prolific, we used the same conditions as in MTurk. The only difference is that we did not restrict recruitment to the country of residence. MTurk participants are concentrated in the U.S., while Prolific participants are dispersed across various countries, although they include more participants from the UK. Therefore, we did not restrict the country of residence of workers in Prolific, as data collected might differ from the original data obtained in Prolific if we restricted the country of residence.

To ensure appropriate consideration of research ethics, informed consent was provided at the time of participant recruitment. Specifically, only those participants who agreed to participate in the survey were administered the questionnaire, after they were informed of the survey content, data handling methods, estimated time required, and amount of compensation. Every participant that answered all the questions was paid in full, regardless of the content of their answers.

Table 2: Results of grouping participants (Number / percentile).

Groups	MTurk	Prolific
G1	122 / 40.7%	13 / 4.3%
G2	0 / 0.0%	0 / 0.0%
G3	34 / 11.3%	122 / 40.7%
G4	143 / 47.7%	164 / 54.7%
Sum	299 / 100.0%	299 / 100.0%

4 EXPERIMENTAL RESULTS

In this section, we discuss the results obtained from our online survey experiment. First, we present the results obtained from classifying the participants into the groups defined in Table 1, including the bias introduced by each screening method on the distribution of the population. Next, we demonstrate how the survey results on security knowledge and security behavior are affected by the screening methods. Finally, we present how the experimental results on the phishing email detection task performance are influenced by the screening methods.

4.1 Classification of participants

We recruited a total of 600 participants, 300 for each of the two crowdsourcing platforms, MTurk and Prolific. We restricted participation to workers with a past task approval rate exceeding 95% on both crowdsourcing platforms. Accordingly, all 600 participants passed the CAPTCHA test at the beginning of the questionnaire.

Table 2 presents the results obtained from classifying the participants into the groups defined in Table 1. Here, both MTurk and Prolific had one participant who did not pass the attention test but passed the IMC, which is more difficult to identify. In this study, these two participants were excluded from the analysis. Hence, the total number of participants was 299 each. First, we can observe that the participants in the two crowdsourcing platforms exhibit significantly different group distributions. For example, in MTurk, the percentage of dishonest participants belonging to Group 1 exceeds 40%, while in Prolific, the percentage is only approximately 4%. The percentage of participants belonging to G3 is also significantly different: 11% in MTurk and 41% in Prolific. The similarity between the two is that no participant was classified as G2. Next, we can infer that the commonality in the experiments of both crowdsourcing platforms is that there were no participants classified as G2. In other words, among the honest participants who passed the completion time and open-ended question screenings, there were no participants who did not pass both the attention check and IMC screenings; however, they were either participants who did not pass the IMC (G3) or participants who passed both (G4).

Most of the participants who were classified as dishonest (G1) did not pass the open-ended question screening. Among the low-quality responses detected by the open-ended question screening, a few responses were copied and pasted from texts available on the Internet. In MTurk, the overall average number of characters in the open-ended responses was 59.8, the average number of characters for dishonest participants (G1) was 32.5, and the average number of characters for honest participants (G3 or G4) was 78.1. Similarly, for Prolific, the overall average was 82.2 characters, while the average for dishonest (G1) and honest (G3 or G4) participants were 25.3 and 84.8, respectively. Both results indicate that dishonest participants tend to have fewer characters in their comments. Interestingly, for the G1 participants, MTurk G1 participants tended to have more characters than their Prolific counterparts. This result is obtained because dishonest participants of Prolific tended to write short sentences such as yes/no, while MTurk dishonest participants copied text from the Internet in several cases. This suggests that dishonest participants in MTurk attempt to circumvent screening in a more sophisticated way.

In this study, open-ended questions were set up to discard the responses of dishonest participants. The fact that the participants who passed the screening with open-ended questions were not classified as G2, which is defined as having a low level of attention, suggests that it may work as well as or better than the attention test in removing careless participants.

A corollary finding, suggested by these results, is the existence of significant differences in participant tendencies across the two crowdsourcing platforms. The observation that there are differences in the participants of the two platforms is consistent with the report by Adam et al. [1]. In our experiment, the percentage of dishonest participants was significantly higher in MTurk. It is interesting to observe such a significant difference, even though we recruited workers with a task approval rate exceeding 95% on both crowdsourcing platforms. The average time taken to complete the entire survey in each of MTurk/Prolific was 7.4/8.9 min, 9.6/12.4 min, and 10.4/13.3 min for Groups 1, 3, and 4, respectively. We can infer that the more honest and attentive the participants were, the more time they took to answer the questions, and in general, the MTurk participants took less time to complete the questions. It is important to test our hypothesis on crowdsourcing platforms with such different characteristics.

The demographic characteristics of the participants in each group are presented in Table 3 and Figure 5. The table presents the breakdown of the number of participants in each group, and the figure illustrates the ratios of participants in each group. First, we can observe that screening alters the demographic distribution of participants, and that the degree of change varies across crowdsourcing platforms. Particularly, in MTurk, the change in demography owing to screening was significant. The results obtained in MTurk indicated that participants who were more likely to be removed by IMC had the following attributes: male, older age group, no college degree, and no IT work experience. This result is mostly consistent with the findings of a previous study [18], i.e., participants who are male and do not have a degree are more likely to be removed by IMC. However, in previous studies [5, 18], the responses of younger participants tended to be more likely to be discarded by IMC, and we obtained different results. In contrast, for Prolific, the changes owing to screening were milder. However, as we will discuss later, the experimental results of the phishing detection performance differed, depending on the screening for Prolific. Therefore, although there appears to be less changes in demography, screening may have affected factors that did not appear in the headcount data.

Table 3: Demographics of participants: MTurk (top) and Prolific (bottom).

Groups	Female/male	Under/above 34 years	With/without college degree	With/without IT work experience
G1	58/64	76/46	116/6	116/6
G3	9/25	21/13	20/14	12/22
G4	51/92	86/57	112/31	86/57
Overall	118/181	183/116	248/51	214/85

Groups	Female/male	Under/above 34 years	With/without college degree	With/without IT work experience
G1	7/6	11/2	4/9	11/2
G3	68/54	95/27	56/66	32/90
G4	77/87	143/21	80/84	39/125
Overall	152/147	249/50	140/159	82/217

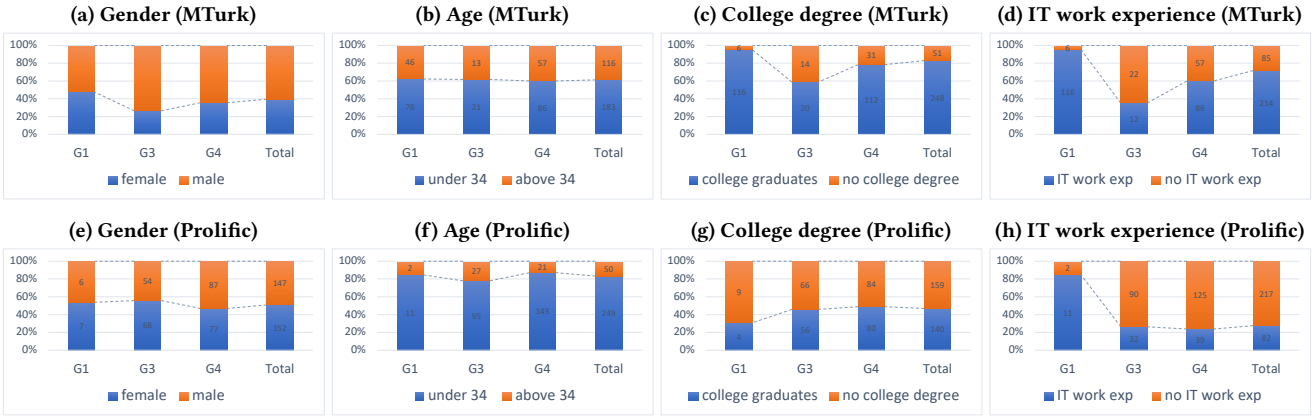

Figure 5: Demographics for each group.

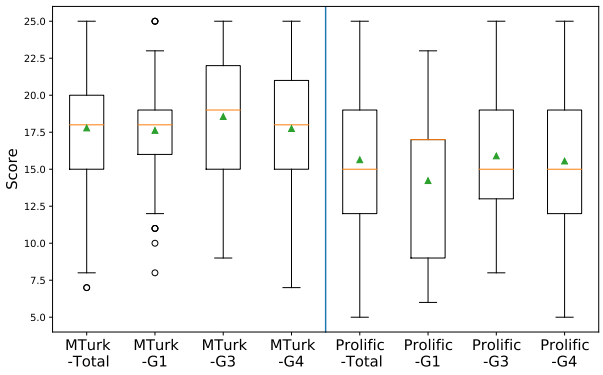

Figure 6: Distributions of security knowledge scores: MTurk (left) and Prolific (right).

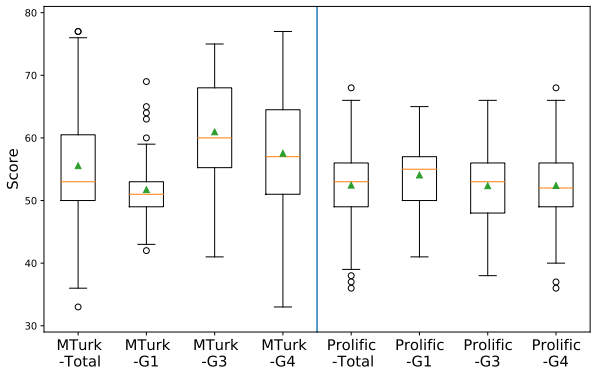

Figure 7: Distributions of SeBIS scores: MTurk (left) and Prolific (right).

4.2 Security knowledge and behavior

We used Part 2 of the questionnaire to assess the impact of screening on an experiment investigating the security knowledge and security behavior of participants. The results obtained from this experiment are presented in Figures 6 and 7. The figures are box plots, where the top and bottom of the boxes represent the first and third quantiles, respectively, and the band inside the box depicts the median. The whiskers of the plot represent the lowest/highest datum within 1.5 IQR of the first/third quantile, where IQR is the difference between the first and third quantiles. In addition, the triangles represent

Table 4: Email task completion time.

Groups	MTurk	Prolific
G1	2.14 min	3.10 min
G3	3.69 min	4.78 min
G4	3.99 min	5.54 min
Overall	3.22 min	5.20 min

mean values. Outliers beyond the whiskers are represented with circle symbols.

In general, we infer that screening influences the results of both experiments that examine the security knowledge and measure the security behavior of participants. We also determine that the degree of impact is greater for MTurk participants. This observation is similar to the trend in the demographics observed in the previous section. In the results of this experiment, we are particularly interested in the difference between Group 3 and Group 4, as the difference between these groups reflects the difference in whether or not they were excluded via the application of IMC[1]. We conducted t-tests on these groups. Regarding security knowledge, there was no significant difference between G3 and G4 for both MTurk and Prolific. In terms of security behavior, the overall score for MTurk was significantly higher for G3 participants than for G4 participants (t-test, $p < 0.05$). This result may seem paradoxical at first glance, in that participants with moderate attention levels (G3) reported more desirable security behavior than those with high attention levels (G4). However, the findings we obtain from G3 have important implications. Careless participants in G3 may fall for phishing sites because of their carelessness, even though they are actually engaging in the correct security behavior. In view of this finding, we can conclude that phishing detection performance is affected not only by security behavior, but also by carelessness. If we remove careless participants from the survey by applying a screening like IMC, we risk losing the useful insights, such as the one described above.

4.3 Evaluation of phishing email detection performance

We used Part 3 of the questionnaire to assess the impact of screening on an experiment measuring the ability of participants to detect phishing. We first present the time required to complete each task in Table 4. It can be observed from the table that in both crowdsourcing platforms, the more honest and attentive participants spent more time on the phishing detection task. Next, we present our results on the detection accuracy in Figures 8 and 9. The legitimate/phishing email detection accuracy is a measure of the percentage of legitimate/phishing emails that were correctly detected as legitimate/phishing. In general, the results obtained from experiments using both crowdsourcing platforms varied, depending on the screening applied. This tendency is more significant than in the survey on security knowledge and security behavior, as seen in the previous section.

[1]Although we were originally interested in the participants in Group 2, who were excluded by applying the attention check, in this experiment, there were no participants in Group 2.

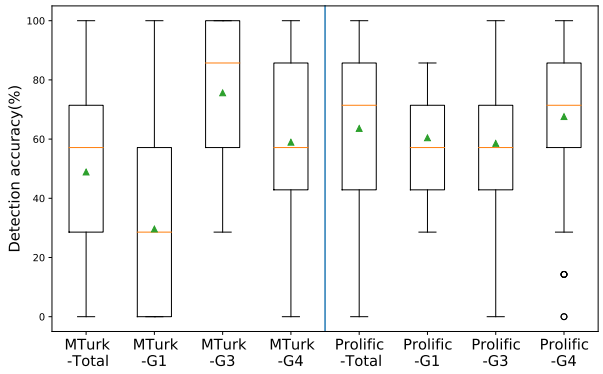

Figure 8: Legitimate email detection accuracy: MTurk (left) and Prolific (right).

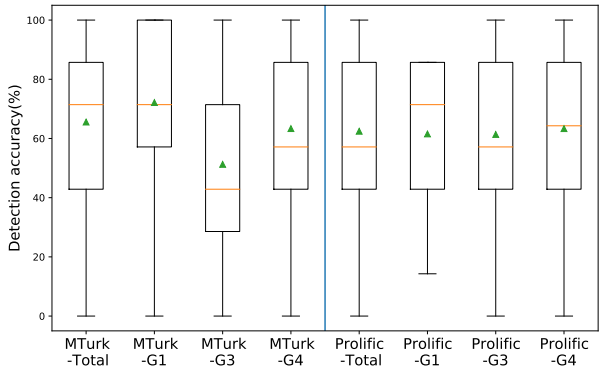

Figure 9: Phishing email detection accuracy: MTurk (left) and Prolific (right).

We can see that the results of dishonest participants in Group 1 deviate from those of honest participants in Groups 3 and 4 in both legitimate and phishing email detection accuracy. In particular, MTurk is unique in that the legitimate phishing email detection rate was low and the phishing email detection rate was high, which is owing to the fact that many participants answered "Yes" to "Is this a phishing email?" in all the tasks. This observation suggests that many of the answers from dishonest participants in Group 1 are noise that should be excluded in the study of user security.

Next, we compared the results of participants with moderate attention levels in Group 3 and those with high attention levels in Group 4. For MTurk, there was a significant difference in the legitimate email and phishing email detection accuracies between the two groups. The results of the t-test present $p < 0.01$ and $p < 0.05$ for legitimate email and phishing email detection accuracies, respectively. Similarly, for Prolific, only the legitimate email detection accuracy was significantly different between the two groups ($p < 0.05$); there was no significant difference in the phishing email detection accuracy ($p = 0.257$). In general, it can be

observed that the results for both crowdsourcing platforms vary by applying screening. It can also be observed that no definite rule exists for the impact of applying screening. In addition, the alterations also vary depending on crowdsourcing platform. As the causes of these changes are not obvious, further insight is required. These non-trivial results may be owing to the fact that attention, a hidden factor that IMC excludes, has a non-trivial correlation with other factors that determine the security behavior of users. Because there is a risk of missing such hidden factors by applying screening too casually, researchers need to carefully consider the impact of such factors on their experiments.

5 DISCUSSION

In this section, we first discuss several challenges in applying screening methods in online surveys. We also provide recommendations for researchers who are planning to apply screening methods. Next, we present the limitations of this work and future research challenges.

5.1 Challenges in applying screening methods

The results from our study indicate that there is a concern that the approach selected by researchers to implement screening methods in their online survey may significantly influence the results of their study. Therefore, researchers should carefully consider the adoption of screening methods according to their research purpose. To help researchers adopt screening methods appropriately, we discuss the challenges of the screening methods.

Worker's reputation

Our results indicate that adopting open-ended questions as a screening method can help identify dishonest participants (Group 1 in Table 1). In particular, despite recruiting participants solely with a past task approval rate exceeding 95%, 40.7% and 4.3% of participants in MTurk and Prolific, respectively, were identified as dishonest. The results suggest that a worker's reputation in a cloud sourcing service (e.g., past task approval rate, the number of tasks approved) does not necessarily reflect the quality of the worker's responses. Researchers need to apply their own screening methods to improve the quality of data.

Automating the analysis of open-ended questions.

While we verified that employing open-ended questions is an effective screening method in discarding dishonest responses, the analysis of open-ended responses requires manual effort, which depends on their skill and experience. Such manual efforts may interfere with large-scale surveys. The combination of natural language processing and machine learning has the potential to automate such analysis, which is left for a future research topic.

Addressing demographic bias.

As this work and previous studies have revealed, the adoption of screening methods may introduce demographic bias, which could significantly impact the results of online survey. In fact, some services such as Prolific Academic and Qualtrics recommend not using attention checks owing to the bias concerns in the participant demographics [30]. Given these observations, we encourage researchers to examine whether biases introduced by screening have influenced their results. If applying a screening method seems to produce a large bias, and if that bias has a significant impact on the results, it implies the need to revise the screening method. The development of a screening method that can minimize the bias on user demography is an open research problem.

Experiments where user attention is an important factor.

In user security research, user attention can be a crucial factor in understanding user perception and behavior. Although IMC can remove participants who are not highly attentive, the removed participants are likely to be potential phishing victims. Therefore, the adoption of IMC poses the risk of removing appropriate participants for security research, especially in phishing surveys where the level of attention of the participants may influence results. Furthermore, because in general, the attentive level of humans easily changes over time, the results of an attention check or IMC in a specific user study do not necessarily reflect the permanent nature of participants [3, 5]. Therefore, we recommend security researchers to carefully interpret the results of an attention check or IMC, especially when user attention can be a factor that influences the results.

5.2 Limitations and future work

Although our study is the first to demonstrate that screening methods may generate biases in user security studies, we identify a few limitations of our study. For future studies, we recommend the following perspectives corresponding to the identified limitations.

Number and position of questions for screening.

We simultaneously implemented five types of screening methods in our questionnaire. Hence, most of the participants would have been directly aware of their explicit types, such as attention checks and IMC. As existing studies also mentioned [15, 17], this awareness may increase the attention level of participants to be more than that of a conventional survey (i.e., questionnaire without explicit types of screening method). In addition, CAPTCHA, attention checks, and IMC were placed in fixed positions in our questionnaire. It is crucial to verify whether the same result as in this study can be obtained by altering the number and position of questions for screening in future work.

Other screening methods.

Although various screening methods have been applied in existing studies, we implemented only five representative types among them in our questionnaire. Other methods [7, 29, 31], which include distribution of scale options, click count, IP address uniqueness, trap questions (e.g., "I was born in the 1700s"), reversed questions (e.g., "I tend to be organized" vs. "I tend to be disorganized"), should also be verified for their effectiveness, and issues should be clarified in the same way as in this study. Furthermore, because there may be advanced bots that evade detection in crowdsourcing services, it is necessary to explore a method that eliminates dishonest responses (especially from advanced bots) with higher accuracy.

6 SUMMARY

The objective of this study was to elucidate the effect of screening methods in online user surveys on the results of security research, especially user security studies on phishing attacks. We conducted an online survey experiment on security knowledge, security behavior, and phishing email detection performance using MTurk and Prolific, two typical crowdsourcing platforms adopted in online

surveys, and obtained the following results. First, we observed that we could effectively screen dishonest participants (Group 1) by applying a screening method based on response completion time and open-ended questions. Next, we focused on the participants who responded honestly. By comparing participants with moderate attention levels (Group 3) and participants with high attention levels (Group 4), we observed that applying the IMC screening method resulted in a bias in the demographics of the participants, as well as a difference in the results of the phishing email detection performance. We also observed that the degree of these differences varies across crowdsourcing platforms, and that the correlation between screening and multiple factors influencing the results is unclear. These observations suggest that caution should be exercised when applying screening methods, such as attention check and IMC, in studies where the extent of user attention has a significant impact on the results. Because user behavior against security threats is largely considered to be correlated with user attention, developing an effective screening method for conducting online surveys on security behavior remains a challenge that must be addressed in future studies. In addition, studying the effectiveness of methods other than the screening methods examined in this study, their effects on the results, and the theories that generate these effects are future challenges.

REFERENCES

[1] Troy L Adams, Yuanxia Li, and Hao Liu. 2020. A Replication of Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research–Sometimes Preferable to Student Groups. *AIS Transactions on Replication Research* 6 (2020).

[2] Amazon. [n.d.]. Amazon Mechanical Turk. https://www.mturk.com/.

[3] Eva Anduiza and Carol Galais. 2017. Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research* 29, 3 (2017), 497–519.

[4] Oshrat Ayalon and Eran Toch. 2019. Evaluating users' perceptions about a system's privacy: differentiating social and institutional aspects. In *Proceedings of the fifteenth symposium on usable privacy and security (SOUPS'19)*.

[5] Adam J Berinsky, Michele F Margolis, and Michael W Sances. 2014. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* 58, 3 (2014), 739–753.

[6] Adam J Berinsky, Michele F Margolis, Michael W Sances, and Christopher Warshaw. 2021. Using screeners to measure respondent attention on self-administered surveys: Which items and how many? *Political Science Research and Methods* 9, 2 (2021), 430–437.

[7] Erin M. Buchanan and John E. Scofield. 2018. Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods* 50, 6 (2018), 2586–2596.

[8] Casey Inez Canfield, Fischhoff Baruch, and Davis Alex. 2016. Quantifying phishing susceptibility for detection and behavior decisions. *Human factors* 58, 8 (2016), 1158–1172.

[9] Sauvik Das, Laura A Dabbish, and Jason I Hong. 2019. A typology of perceived triggers for end-user security and privacy behaviors. In *Proceedings of the fifteenth symposium on usable privacy and security (SOUPS'19)*.

[10] David L Dickinson and David M McEvoy. 2020. Further from the Truth: The Impact of In-Person, Online, and mTurk on Dishonest Behavior. *IZA Discussion Paper* 13686 (2020).

[11] Julie S. Downs, Mandy B. Holbrook, and Lorrie Faith Cranor. 2006. Decision Strategies and Susceptibility to Phishing. In *Proceedings of the Second Symposium on Usable Privacy and Security*. Association for Computing Machinery, 79–90. https://doi.org/10.1145/1143120.1143131

[12] Marc Dupuis, Emanuele Meier, and Félix Cuneo. 2019. Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavior Research Methods* 51, 5 (2019), 2228–2237.

[13] Serge Egelman and Eyal Peer. 2015. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*.

[14] Florian M. Farke, Lennart Lorenz, Theodor Schnitzler, Philipp Markert, and Markus Dürmuth. 2020. "You still use the password after all" – Exploring FIDO2 Security Keys in a Small Company. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, 19–35. https://www.usenix.org/conference/soups2020/presentation/farke

[15] David Hauser and Norbert Schwarz. 2015. It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks. *SAGE Open* 5 (03 2015). https://doi.org/10.1177/2158244015584617

[16] David J Hauser and Norbert Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav Res Methods.* 48, 1 (2016), 400–407.

[17] Michael S. Jones, Lisa A. House, and Zhifeng Gao. 2015. Respondent Screening and Revealed Preference Axioms: Testing Quarantining Methods for Enhanced Data Quality in Web Panel Surveys. *Public Opinion Quarterly* 79, 3 (06 2015), 687–709. https://doi.org/10.1093/poq/nfv015 arXiv:https://academic.oup.com/poq/article-pdf/79/3/687/5407002/nfv015.pdf

[18] Adam Kapelner and Dana Chandler. 2010. Preventing satisficing in online surveys: A "Kapcha" to Ensure Higher Quality Data. *Proceedings of the CrowdConf'10* (2010).

[19] Laura Lascau, Sandy Gould, Anna Cox, Elizaveta Karmannaya, and Duncan Brumby. 2019. Monotasking or Multitasking: Designing for Crowdworkers' Preferences. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

[20] Aaron J Moss and Leib Litman. 2018. After the bot scare: Understanding what' s been happening with data collection on MTurk and how to stop it. *Retrieved February* 4 (2018), 2019.

[21] U.S. Department of Labor; Wage and Hour Division. [n.d.]. State Minimum Wage Laws. https://www.dol.gov/agencies/whd/minimum-wage/state.

[22] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology* 45, 4 (2009), 867–872.

[23] Jonas Oppenlaender, Kristy Milland, Aku Visuri, Panos Ipeirotis, and Simo Hosio. 2020. Creativity on Paid Crowdsourcing Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[24] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

[25] Prolific. [n.d.]. Prolific Academic. https://www.prolific.co/.

[26] Qualtrics. [n.d.]. Qualtrics. https://www.qualtrics.com/.

[27] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. 2019. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *2019 IEEE Symposium on Security and Privacy (SP)*. 1326–1343. https://doi.org/10.1109/SP.2019.00014

[28] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the 2010 SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*.

[29] Andie Storozuk, Marilyn Ashley, Véronic Delage, and Erin A Maloney. 2020. Got Bots? Practical Recommendations to Protect Online Survey Data from Bot Attacks. *The Quantitative Methods for Psychology* 16, 5 (2020), 472–481.

[30] Dave Vannette. 2017. 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/.

[31] Christina Yarrish, Laurie Groshon, J Mitchell, Ashlyn Appelbaum, Samantha Klock, Taylor Winternitz, and Dara G Friedman-Wheeler. 2019. Finding the signal in the noise: Minimizing responses from bots and inattentive humans in online research. *The Behavior Therapist* 42, 7 (2019), 235–242.

APPENDIX A: QUESTIONNAIRE

Part 1

(1) Which gender do you identify with the most?
- ○ Female
- ○ Male
- ○ Other
- ○ Prefer not to answer

(2) What is your age range?
- ○ 18–24
- ○ 25–34
- ○ 35–44
- ○ 45–54
- ○ 55–64
- ○ 65 or more
- ○ Prefer not to answer

(3) What is the highest level of school you have completed or degree you have earned?
- ○ Less than high school
- ○ High school or equivalent
- ○ College or associate degree
- ○ Bachelor's degree
- ○ Master's degree
- ○ Doctoral degree
- ○ Professional degree
- ○ Other
- ○ Prefer not to answer

(4) Have you held a job in computer science, information technology, or a related field?
- ○ Yes
- ○ No

(5) How many computer/mobile devices do you use? (desktop PCs, laptop PCs, tablets, or smartphones)

(6) Which device are you using for answering this survey now?
- ○ Desktop PC
- ○ Laptop PC
- ○ Tablet
- ○ Smartphone
- ○ Other

(7) How many hours per day do you use your smartphone on average?
- ○ Less than 1h
- ○ More than 1h but less than 2h
- ○ More than 2h but less than 3h
- ○ More than 3h but less than 4h
- ○ More than 4h but less than 5h
- ○ More than 5h but less than 6h
- ○ More than 6h

(8) Which social media do you use? (multiple choices allowed)
- □ Facebook
- □ Twitter
- □ Instagram
- □ YouTube
- □ LinkedIn
- □ WhatsApp
- □ Snapchat
- □ Other
- □ I do not use social media

(9) Do you use a password manager?
- ○ Yes
- ○ No
- ○ I do not know

(10) Do you use two-factor authentication for any of your online accounts?
- ○ Yes
- ○ No
- ○ I do not know

(11) Do you think you are in control of your personal information?
- ○ Yes
- ○ No

(12) On the basis of your answer to Q11, why do you think that[2]?

(13) Have you ever received anti-phishing training?
Phishing: Online fraud that acquires sensitive information primarily by masquerading as a legitimate business or reputable person.
- ○ Yes
- ○ No
- ○ I do not know

(14) 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].
- ○ Yes
- ○ No
- ○ I do not know

(15) Have you ever been deceived by a phishing email?
Being deceived by a phishing email means that you visited a website linked in the phishing email or opened a file attached to the phishing email, regardless of whether it was directly damaged or impacted you.
- ○ Yes
- ○ No
- ○ I do not know

Part 2

(1) How familiar are you with the following network- or security-related items?
Please choose from the following scale: no understanding - little understanding - some understanding - good understanding - full understanding.

	None	Little	Some	Good	Full
IP address	○	○	○	○	○
Malware	○	○	○	○	○
Cookie	○	○	○	○	○
SSL/TLS	○	○	○	○	○
VPN	○	○	○	○	○

(2) Please indicate how often you have done the following descriptions on the following scale:
Never - Rarely - Sometimes - Often - Always.
- I set my computer screen to automatically lock if I do not use it for a prolonged period of time.
- I use a password/passcode to unlock my laptop or tablet.
- I manually lock my computer screen when I step away from it.
- I use a PIN or passcode to unlock my mobile phone.
- I do not change my passwords, unless I have to[4].

[2]This question was set as a screening method (Open-ended question).
[3]This question was set as a screening method (IMC).
[4]These questions are reverse-scored questions (Always: 1-point, Often: 2-points, Sometimes: 3-points, Rarely: 4-points, Never: 5-points).

- I use different passwords for different accounts that I have.
- When I create a new online account, I try to use a password that goes beyond the site's minimum requirements.
- I do not include special characters in my password if it's not required[4].
- When someone sends me a link, I open it without first verifying where it goes[4].
- I know what website I'm visiting based on its look and feel, rather than by looking at the URL bar[4].
- I submit information to websites without first verifying that it will be sent securely (e.g., SSL, "https://", a lock icon)[4].
- Please select "Never" for this question[5].
- When browsing websites, I mouseover links to see where they go, before clicking them[4].
- If I discover a security problem, I continue what I was doing because I assume someone else will fix it.
- When I'm prompted about a software update, I install it right away.
- I try to make sure that the programs I use are up-to-date.
- I verify that my anti-virus software has been regularly updating itself.

Part 3

(1) From here, we will evaluate a series of emails to determine whether or not they are phishing emails. Phishing is an online fraud that acquires sensitive information primarily by masquerading as a legitimate business or reputable person. There will be 14 emails (phishing emails will be included). You will review Kelly Harmon's email. Please imagine that you are receiving these emails and answer the questions to the best of your ability without an Internet search. To help you answer the questions, some information about Kelly is listed below. You will be able to refer to this information while answering the questions.

- Is this a phishing email?
  <A screenshot of each email>
  - Yes
  - No

  Information about Kelly:
  Name: Kelly Harmon
  Company name: Soma Corporation
  In-house IT-related contact: IT Help Desk(helpdesk@soma.com)
  Eternal services in use: Google Voice, eBay, Apple(iTunes), Bank of America, Capital One, LinkedIn, Netflix

[5]This question was set as a screening method (Attention check).

APPENDIX B: SUMMARY OF EMAILS.

Table 5: Summary of emails used for the experiments [8].

	Subject	Sender
Legitimate	Important Phishing Notice - Please Read	Mary Ann Bane <mabane@soma.com>
Legitimate	New voicemail from (724) 970-8435 at 12:27 PM	Google Voice <voice-noreply@google.com>
Legitimate	Important - eBay Password Reset Required	eBay <eBay@reply1.ebay.com>
Legitimate	Your credit card is about to expire	Netflix <info@mailer.netflix.com>
Legitimate	eBay Reset Your Password	eBay <ebay@ebay.com>
Legitimate	Your receipt No.130086326136	iTunes Store <do-no-reply@itunes.com>
Legitimate	Scanned Document From PRINT4.SOMA.COM	PRINT4-SOMA <print4@soma.com>
Phishing	Your Apple ID was disabled	Apple <accounts@apple.com>
Phishing	Customer Alert	Capital One <capitalone@gmail.com>
Phishing	Double Frequent Flyer Miles!	Customer Appreciation <cust@boa.com>
Phishing	Your Email Account	IT Help Desk <helpdesk@soma.com>
Phishing	Invitation to connect on Linkedin	Linkedin <member@linkedin.com>
Phishing	Cyber Security Awareness Month: Take Security 101	Mary Ann Bane <mabane@soma.com>
Phishing	Password will expire in 4 days	IT Help Desk <helpdesk@soma.com>