Personalized Phishing Interventions: Assessing the Efficacy of Phishing Training based on User Knowledge, Ability, and Awareness

Victor Carles
Supervised by Prof. Verena Zimmermann and Lorin Schöni
M.Sc. Cyber Security, EPFL-ETHZ
ETHZ Security, Privacy & Society Lab

June 25, 2023

Abstract

The aim of this study was to evaluate the efficacy of personalized and human-centered interventions in assisting users in identifying phishing emails. While previous research has identified various human factors that contribute to susceptibility to phishing attacks, limited knowledge exists on the most effective interventions for different user categories. This study seeked to determine the effectiveness of diverse interventions based on user groups categorized by prior knowledge on phishing, frequency of email use and email rate as influencing factors in the literature. For low-knowledge users, education modules proved to be highly effective. Training interventions demonstrated positive impacts on users with low to medium knowledge levels, and reminders were shown to be effective across all user groups, including those with high knowledge. Additionally, the study explored potential correlations between personality traits, demographic factors, and the effectiveness of the interventions. While limited significance was found for demographic factors overall, age influenced the impact of interventions for low-knowledge users, and education and IT background were relevant for personalization. Trust emerged as an important trait to consider when designing interventions, while sociability was identified as a significant factor influencing susceptibility to phishing. Based on these results, the study demonstrated that tailoring interventions based on knowledge, demographic factors, and individual characteristics can enhance their effectiveness in improving individual's ability to recognize phishing. The findings from this study could inform the further development of a personalized anti-phishing tool, potentially leading to the creation of a fully functional Gmail plug-in or an application that incorporates the identified effective human-centered interventions.

1 Introduction

Phishing attacks continue to pose a significant threat to both individuals and organizations, despite increased awareness and improved security measures. Although there are technical interventions available to combat phishing, it is important to acknowledge their limitations. Attackers continually evolve their tactics, finding ways to bypass technical measures and exploit human vulnerabilities. Humans are often the weakest link in the security chain, susceptible to social engineering techniques employed in phishing attacks. Focusing on humans is therefore essential because it recognizes the need to address the psychological and behavioral aspects of phishing. By understanding human vulnerabilities, personalized interventions can be developed. Classifying users into categories based on their knowledge, behaviors, or susceptibility to phishing allows for tailored education and training programs. Individuals can then receive targeted protection based on their specific needs. The purpose of designing a personalized intervention tool is to effectively address the weaknesses of each individual and enhance

their resilience against phishing attacks. The goal is to empower individuals to develop their own protection mechanisms, tailored to their specific vulnerabilities, so that they can defend themselves against even the most sophisticated phishing attacks that may circumvent technical safeguards.

Numerous research efforts have attempted to identify the complex array of factors that make users more or less susceptible to phishing attacks. It can be even more difficult to determine how to mitigate these factors so that users can build up resistance towards phishing. In the first part of the study, my goal was to identify, based on literature, potential human factors that could be the most influential in the susceptibility of a user to fall for a phishing email. These factors were used to classify users in order to provide some sort of high-level personalization in the interventions proposed to them. The interventions had the objective of not only helping the users significantly increase their ability to detect phishing, but also be adapted to the category of the users determined prior to the exposure, which could serve the future design of a personalized phishing intervention tool. The second part of the study was then to assess how effective the classification I have structured was and how impactful the interventions were. This part was divided into two segments: the implementation of the materials used during the investigation, and the study per se with the analysis of the data gathered. In the first segment, I implemented a web app that served as a vector for the study, as well as all the material required for the data collection, such as the questionnaires to assess the user's category, the logic behind the classification and the Figma interface mocking a Gmail mailbox. In the second segment, the study was conducted with Prolific. Participants were categorized in one of three groups based on a score derived from a pre-intervention questionnaire on Qualtrics. They were then presented to a Gmail mockup while receiving personalized interventions specific to their group's category. Following this interaction, participants were asked to answer the same set of questions as for the first questionnaire, with the newly acquired knowledge. The study's final phase consisted in analyzing the results, which demonstrated the effectiveness of the interventions and explored possible correlations between human factors and their efficacy.

2 Literature Review: Human factors for phishing susceptibility and appropriate interventions

The initial phase of my study involved conducting a comprehensive literature review to identify the key human factors that significantly influence an individual's susceptibility to falling for a phishing email. In the following, the main findings with regards to the relevant human factors identified in the literature, that is knowledge, frequency of email use, email rate, demographic information and personality traits, will be described. Afterwards, the implications for the study will be summarized.

2.1 Degree of knowledge

Various literature (cf Baral and Arachchilage [2019], Alnajim and Munro [2009], Dou et al. [2017], Desolda et al. [2021]) indicates that knowledge about phishing is a crucial factor in avoiding falling for phishing scams. Unaware users are more susceptible, while experienced users develop techniques to recognize and avoid phishing attempts. However, research (cf Das and Camp [2022], Wang and Rao [2016]) demonstrates that even with knowledge and experience, overconfidence can still lead to falling for phishing scams again. Thus, users must stay updated and alert, and continuous training can improve their ability to recognize evolving phishing techniques, cf Sheng et al. [2010]. For the purpose of the study, I divided phishing knowledge in two segments:

- The phishing theory knowledge, which encompasses a range of expertise, including understanding various techniques to identify phishing attempts, general awareness of what phishing entails, and how to effectively respond to it. In particular, based on research and widely recognized techniques that have been documented and defined over time on the internet (referencing Abdillah et al. [2022], phi [a,b,c]), I have classified phishing theory knowledge into five key components:

  - Comprehensive Understanding: This encompasses knowledge about phishing, its associated risks, and common characteristics found in phishing attacks. It involves recognizing typical patterns and indicators of phishing attempts.
  - Suspicious Sender Analysis: This component involves the ability to evaluate the legitimacy of a sender's email address. It includes techniques to identify suspicious or fraudulent sender information that may indicate a phishing attempt.
  - Persuasive Tactics Recognition: Recognizing the persuasive tactics employed by phishers is crucial. This entails being able to identify and analyze tactics such as scarcity, authority, and reciprocity that phishers commonly employ. Additionally, it involves detecting suspicious content elements, such as misspellings or outdated logos, which may indicate a phishing email.
  - Suspicious Attachment Analysis: Having the capacity to analyze URLs embedded in emails is essential. This involves examining and differentiating between legitimate and malicious URLs, understanding the structure and components of URLs, and identifying signs of potential phishing in the URL. This also includes the examination of attached files that may pose a potential risk, specifically by considering their nature as indicated by the file extension.
  - Appropriate Response: Knowing how to respond when encountering a phishing email is vital. This component includes reporting the phishing incident and taking measures to protect personal information and devices such as changing compromised credentials.

- The practical knowledge, which refers to the ability of a user to effectively differentiate between a phishing email and a legitimate one. While this knowledge can benefit from an understanding of phishing theory, the actual process of distinguishing emails is distinct from mere theoretical knowledge. It requires real-time application, critical analysis of email elements, intuition, and learning from experience, going beyond the theoretical understanding of phishing.

Currently, there is no specific intervention tailored to different levels of knowledge. To address this, I have incorporated findings from relevant literature along with my own assumptions in order to devise personalized interventions. Previous studies (cf. Canfield and Fischhoff [2018a], Franz et al. [2021], Canfield and Fischhoff [2018b], Sumner and Yuan [2019]) have categorized the main interventions as education, training, and reminders. Due to time constraints and implementation complexity, other interventions were not considered. Consequently, I have chosen to focus on these three types of interventions and explore their potential for personalization across different user categories. I have formulated a hypothesis based on my assumptions and the observed effectiveness of interventions in the aforementioned studies, leading to the following division of interventions:

- In the case of users with low phishing knowledge, the most effective intervention would involve an educational module that offers entertaining lessons and tips on identifying phishing emails. This approach is logically justified by the fact that these users lack the basic knowledge required to comprehend phishing attacks. Therefore, educating them is the most straightforward solution. These users may lack the time or inclination to actively learn about phishing, hence emphasizing the importance of making the educational process enjoyable and engaging. However, it is important to note that this user category should also benefit from interventions designed for higher knowledge levels, such as training and reminders.

- For users with moderate knowledge who may recognize some obvious indicators of phishing but are still vulnerable, a training module with interactive quizzes could prove to be highly effective. This approach would enable them to learn techniques they may be unaware of and enhance their ability to identify phishing attempts by exposing them to real-life scenarios. Additionally, these users would also benefit from reminders.

- Users with high knowledge possess a better understanding of common tactics employed by phishers, but they may still be susceptible if they become complacent. Therefore, alerts and reminders would be valuable in consistently keeping them vigilant against potential phishing emails.

2.2 Frequency of email use and email rate

Two other important factors I considered in my study were the frequency of email use and the email rate. Sarno and Neider [2022] demonstrated that higher email rate leads to cognitive overload which, associated with phishing prevalence, increases the susceptibility of falling for the attacks. Conversely, frequency of use is related to usage experience which was demonstrated in the `2.1` section to be a significant factor. Users with a low frequency of use may be less exposed to phishing attempts but may also be more susceptible to falling for them due to their lack of experience. Users with a high frequency of use, such as multiple times a day, would be the most susceptible to falling for phishing attempts as they are statistically more exposed. However, they might also have sufficient knowledge to recognize phishing emails. Regarding the email rate, a user with fewer emails will potentially be less exposed than a user with a higher rate if the phishing prevalence is low.

The inclusion of these two factors in the study for classifying users between knowledge categories was based on the understanding that knowledge remains the key determinant in assessing a user's susceptibility to phishing attempts, regardless of their frequency of use and email rate. By considering these factors alongside knowledge, the study aimed to establish a more comprehensive classification system. For instance, if a user falls between the medium and high knowledge categories, their classification would depend on the combination of frequency of use and email rate. If the frequency of use and email rate are both relatively low, it reduces the risk of falling for phishing attempts, leading to their classification in the high knowledge category. On the other hand, a user between the medium and high knowledge categories with not a low email rate or not a low frequency of email use would be at greater risk and thus classified as medium knowledge. In such cases, training interventions would be beneficial for enhancing their awareness and defenses against phishing attacks.

2.3 Demographic factors

Demographic factors such as age, gender, origin, and education have been extensively studied as potential predictors of susceptibility to phishing attacks. Regarding age, some studies (cf Kumaraguru et al. [2009]) found that younger individuals, particularly those in the 18-25 age range, were more vulnerable to phishing attacks due to lack of experience and sometimes overconfidence. Conversely, other studies, such as Grilli et al. [2020], showed that older adults were more susceptible to falling for phishing emails due to reduced ability to distinguish genuine emails from phishing emails. The varying susceptibility based on age can be attributed to different tactics and elements used in phishing attacks targeting different age groups. Gender has also been explored as a factor in phishing susceptibility. However, the results have been inconsistent, hence I have excluded this factor from the measurements as I deemed it not relevant.

Education plays a crucial role in users' ability to recognize and avoid phishing attempts, according to literature (cf Tornblad et al. [2021]). Higher quality education, especially in IT or computer-related fields, has been found to increase the likelihood of recognizing phishing emails. However, some studies, such as Liu and Zhang [2020] suggest that more education does not necessarily lead to reduced phishing victimization. In other words, being educated does not replace a specialized training program for phishing prevention and detection. Furthermore, Tornblad et al. [2021] suggested that employees in technical jobs were shown to equal non-technical employees in their ability to discriminate between phishing and legitimate emails, even after receiving anti-phishing training.

Overall, the results of the studies on demographic factors have been inconsistent and sometimes contradictory, this is why I have decided to use these factors as measurement variables in my study in order to see if the results corroborate with what I found in the literature.

2.4 Personality traits

Personality traits have been identified as potential factors that can influence an individual's susceptibility to phishing attacks. Several studies (cf López-Aguilar and Solanas [2021], Cho et al. [2016], Halevi et al. [2013], Ge et al. [2021]) have explored the correlation between personality traits and susceptibility to phishing. The five big personality traits commonly studied are openness, agreeableness, conscientiousness, extraversion, and neuroticism.

Studies have found a positive correlation between openness and phishing susceptibility. Individuals with higher openness scores are more likely to respond to phishing emails and are at a greater risk of falling for such attacks. Agreeableness has been positively correlated with phishing susceptibility. Individuals high in agreeableness are more likely to agree to email requests, click on links, and disclose their information. The effects of conscientiousness on phishing susceptibility are inconsistent. Some studies suggest that conscientious individuals may be more likely to click on links in phishing emails, driven by a desire to perform optimally. High extraversion can increase the risk of falling for phishing attacks as individuals with this trait may be more prone to taking risks and engaging in dangerous online behaviors. The effect of neuroticism on phishing susceptibility is mixed. Some studies indicate that individuals with high neuroticism scores are more likely to click on links in phishing emails.

Due to the difficulty in accurately determining an individual's personality traits, I have only incorporated this as a control variable, focusing on the correlation between certain aspects of the big five personality traits and specific interventions.

3 Methodology and Implementation

To enhance clarity, participants categorized in the "low knowledge" category will be referred to as Group 0, participants classified in the "medium knowledge" category will be referred to as Group 1, and participants categorized in the "high knowledge" category will be referred to as Group 2. This grouping system will facilitate better understanding and organization in the subsequent discussions. As shown in figure 1, the study aimed to evaluate the effectiveness of three types of interventions: education for Group 0, training for Group 0 and 1, and reminders for all groups. The expectation were that these interventions would lead to statistically significant improvements in the users' ability to recognize phishing and adopt an adequate behavior. Additionally, correlations between these improvements and other parameters such as demographic factors and personality traits were measured.

The study was divided into three parts: firstly, determining the participant's category; secondly, immersing the participant in a realistic email usage scenario while exposing them to the interventions; and finally, evaluating the participant's improvement after being exposed to the interventions.

3.1 Questionnaires and scoring system for classification

In order to determine the category of the participant, it was required to ask some questions to assess the knowledge of the participant. This was done through a questionnaire set-up with Qualtrics, where the participant's answers were contributing to a score which was decisional for the category. Reader can find the full questionnaire here. Below is a summary of the questions found in the pre-intervention questionnaire, along with their corresponding point values and how they were utilized to determine the category:

- Frequency of email use and email rate: as seen in the first section, these are important factors in phishing susceptibility and were used for participants with scores in-between categories as an influencing parameter.
- Previous security training experience on phishing: individuals with previous training should have

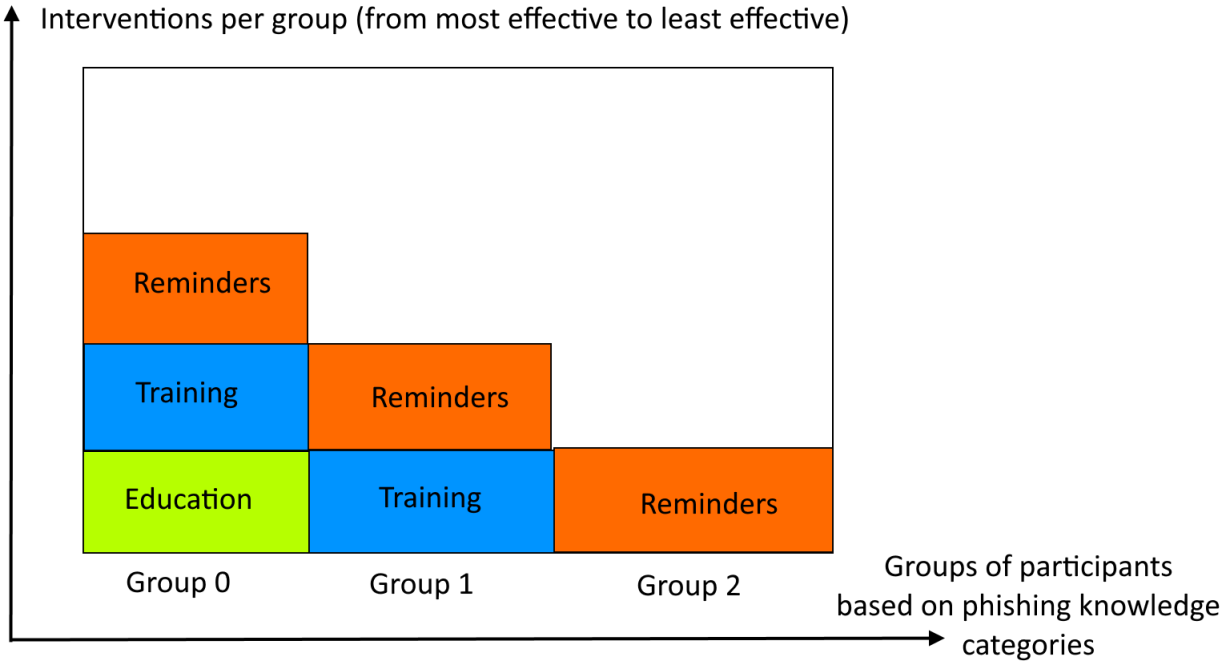

Figure 1: Interventions per group

a better understanding of phishing techniques and concepts, making training experience a more objective indicator. However, without detailed information about the nature and quality of the training, it is challenging to assess its effectiveness. This was included as a bonus for the scoring. The following bonus points were attributed: "Yes, once": 1 point, "Yes, more than once": 2 points.

- Self-report on phishing knowledge: while a user's self-report of phishing knowledge may not always accurately reflect their actual knowledge, it provided an indicator of their perception of their own ability to detect phishing. This was also included as a bonus for the scoring. The self-report consisted in assessing their knowledge on phishing, their ability to detect phishing emails and their level of alertness to phishing attacks. The following bonus points were attributed: "medium": 0.25 points, "rather high": 0.5 points, "very high": 1 point.

- Online security behavior assessment: The susceptibility to falling for phishing attacks can be reduced by demonstrating good online security behavior. The assessment of online security behavior played a significant role in determining the overall score. In the online security form, participants were awarded 0.5 points for providing a strong and correct answer, 0.3 points for a weak answer leaning towards the correct response, 0.1 points for a neutral answer, and 0 points for an opposing answer. For example, if participants were expected to strongly agree with a statement such as "it's risky to open an email attachment from an unknown sender," their score would decrease from 0.5 points to 0 points as their agreement decreased from "Strongly agree" to "Strongly disagree." The online security form contributed a total of 5.5 points towards the overall score.

- Phishing theory knowledge quizz: as seen in the `2.1` section, the phishing theory knowledge was divided in five segments. Comprehensive Understanding questions counted for 12 points in total, which was the biggest weight as the overall knowledge was seen as the most important. Suspicious Sender Analysis questions counted for 6 points, Persuasive Tactics Recognition questions counted for 8 points, Suspicious Attachment Analysis questions counted for 4 points, and Appropriate Response questions counted also for 4 points. The theory knowledge questions yielded a total of 34 points. Summed to the online security behavior assessment, the total phishing theory score

was 39.5.

- Phishing practice knowledge quizz: The participants were exposed to five email examples, similar to figure 2, andwere asked to determine whether each email was a phishing email or not. These emails were created based on real phishing emails, sourced from popular datasets or the my personal spam mailbox, and were designed using Figma. All five example emails can be accessed here. For each correct answer, participants were awarded 2 points, resulting in a total phishing practice score of 10 points. The total score a participant could obtain for both theory and practice, including the bonus points, was 49.5.

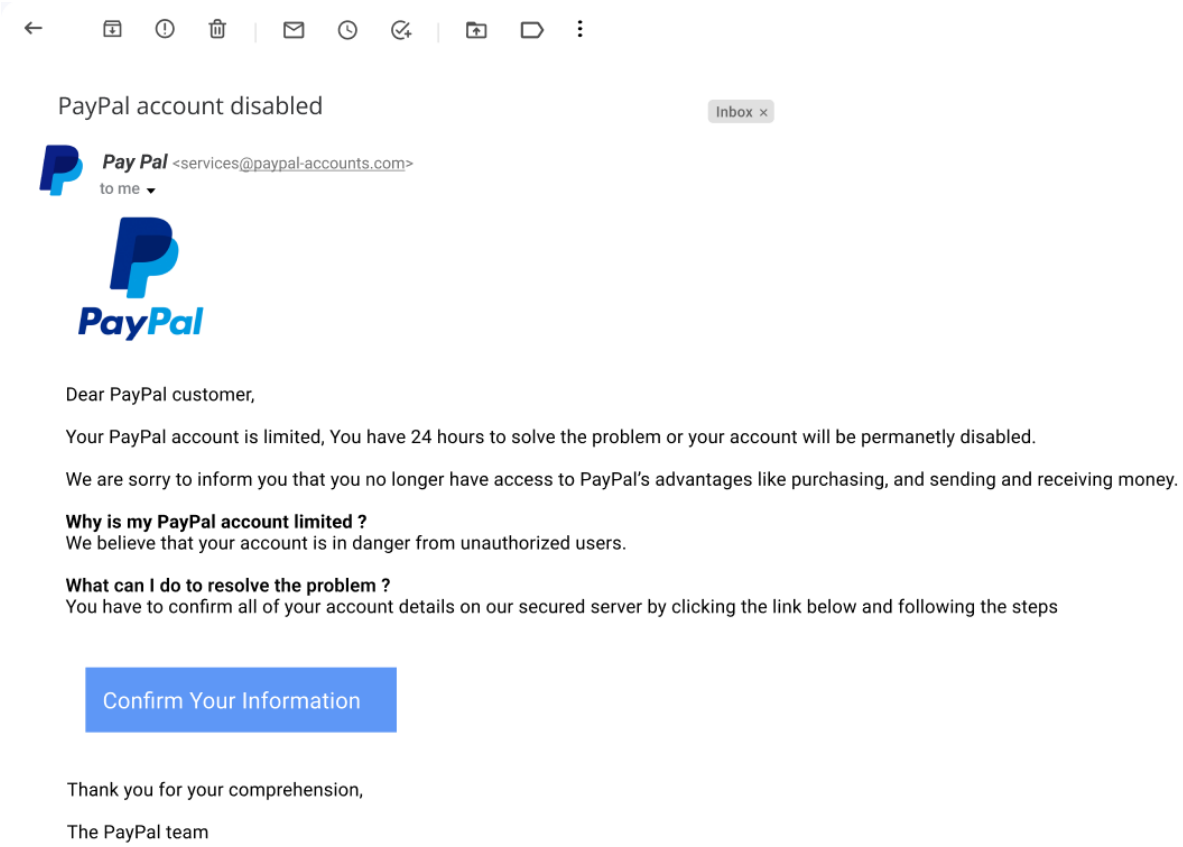

Figure 2: Example of phishing email used for the practice knowledge quizz

The classification was initially layered as follow: a first threshold was set at 13 for the phishing theory score, i.e participants with a total phishing theory score lower than 13 (which was approximately 1/3 of the total number of points they could obtain) were automatically classified in Group 0. Then, the total score would be computed (phishing theory score + phishing practice score), as well as the bonuses (self-report and previous security training experience). A second threshold was evaluated if the total score was less than 25 (0.5 of total score). In this case, the participant were classified in Group 0. Conversely, the third threshold considered that a participant with more than 37 points (approximately 2/3 of total score) was knowledgeable enough to be classified in Group 2. For in-between participants, the frequency of use and email rate were used to determine the category. If participants had low frequency of use and low email rate, they were considered "low risk" and were classified in Group 2. In any other case, the participants were classified in Group 1, as they would still be at risk and require training.

In order to validate this scoring for classification, a first run of the study was done on 10 participants. The results demonstrated that none were classified in Group 0. In order for the study to be un-biased and as broad as possible, it was required that the 100 participants had a distribution of approximately

33% in each group. On the other hand, it was worth noting that participants were gathered from Prolific and the majority of them had already been exposed to security training and phishing-related studies, and self-reported a rather high to very high general knowledge of phishing. Given this information, I voluntarily didn't want to drastically change my scoring system to fit the participants as I still wanted it to be realistic of the general population. Using pandas in python, I computed the mean and standard-deviation for phishing theory knowledge, phishing practice knowledge and total phishing knowledge. Based on new thresholds, computed as in figure 3, I changed the scoring system to set the first threshold at 20.5, second at 27.5 and third at 39.5. This scoring system was the final one that was used during the full study on 100 participants.

To validate the scoring system for classification, an initial study was conducted involving 10 participants. The results revealed that none of the participants were classified in Group 0. To ensure an unbiased and comprehensive study, it was necessary to have a distribution of approximately 33% in each group among the 100 participants. However, it should be noted that the participants were recruited from Prolific, and the majority of them had prior exposure to security training and phishing-related studies. Additionally, they self-reported a relatively high to very high level of general knowledge about phishing. Considering these factors, I intentionally chose not to make drastic changes to the scoring system to accommodate the participants. It was important to maintain realism and reflect the knowledge levels of the general population. As shown in Figure 3, I calculated the mean and standard deviation for phishing theory knowledge, phishing practice knowledge, and total phishing knowledge. The revised scoring system incorporated these new thresholds, with the first threshold set at 20.5, the second at 27.5, and the third at 39.5. This final scoring system was implemented and utilized during the full study involving 100 participants.

```
In [52]: def compute_mean_std(scores):
             mean_score = np.mean(scores)
             std_score = np.std(scores)
             return mean_score, std_score

         def compute_thresholds(mean, std):
             threshold1 = mean - std
             threshold2 = mean + std
             return threshold1, threshold2
```

Figure 3: Functions to compute mean, standard deviations and new thresholds given pandas dataframes

Once the participants had been exposed to the interventions accordingly to their given group, they were redirected to a post-intervention questionnaire which assessed again their score. The full questionnaire can be found here. In order to not introduce any bias in the process, the exact same questions as for the pre-intervention questionnaire were displayed to the participants, in a randomized order. In addition, some personal questions were asked to gather data useful for measurement, as seen in the first section, such as questions on demographics and personality traits. IUIPC and SE13 questions were also asked to the participants. IUIPC (Internet Users' Information Privacy Concerns) is commonly used to assess individuals' concerns and attitudes towards the privacy of their personal information when using the internet. The IUIPC questionnaire typically consists of items or questions that evaluate individuals' perceptions and worries regarding the collection, use, and dissemination of their personal information online. It aims to understand users' privacy concerns and their level of sensitivity towards sharing personal data in online environments. In the context of the study, the answers collected from the IUIPC served as control variables to establish the correlation between privacy concerns and phishing susceptibility, as well as measuring how well a participant with no privacy concern would perform compared to a concerned participant. On the other hand, SE13 questions are designed to assess an individual's knowledge and understanding of general security practices and principles, including those related to phishing attacks. The SE13 questionnaire consists of 13 multiple-choice questions covering various aspects of cybersecurity, such as password security, email practices, social engineering, and recognizing phishing attempts. Once the new score was computed, the participants were redirected to Prolific to finish their participation.

3.2 Gmail mockup

During the study, participants engaged with a Gmail mockup created using Figma, a cloud-based design and prototyping tool known for its collaborative interface design capabilities. To construct the mockup, I utilized the "FREE Gmail Mockup 2023 template!" by Scarlett Rivera, which can be found here, and added my own components. The purpose was to provide participants with a realistic email interaction experience. The full prototype used during the study can be found here. The mockup was meant as a background task designed to be interactive and enjoyable, resembling regular mailbox usage. Figure 4 illustrates the scenario presented to participants. The participants interacted with a variety of fabricated emails for 10 to 15 minutes, including meeting invitations, informative emails containing credentials or guidelines PDFs, fun facts, support tickets, and greeting emails. These emails, designed by myself, incorporated clickable elements such as downloadable files and hyperlinks. It is important to note that these interactions were solely for the purpose of the study and did not access the participant's private computer in any way nor redirect them to external websites. The background task aimed to expose participants to potential vulnerabilities, such as downloading malicious files or being redirected to phishing websites, but everything displayed was meant as a simulation.

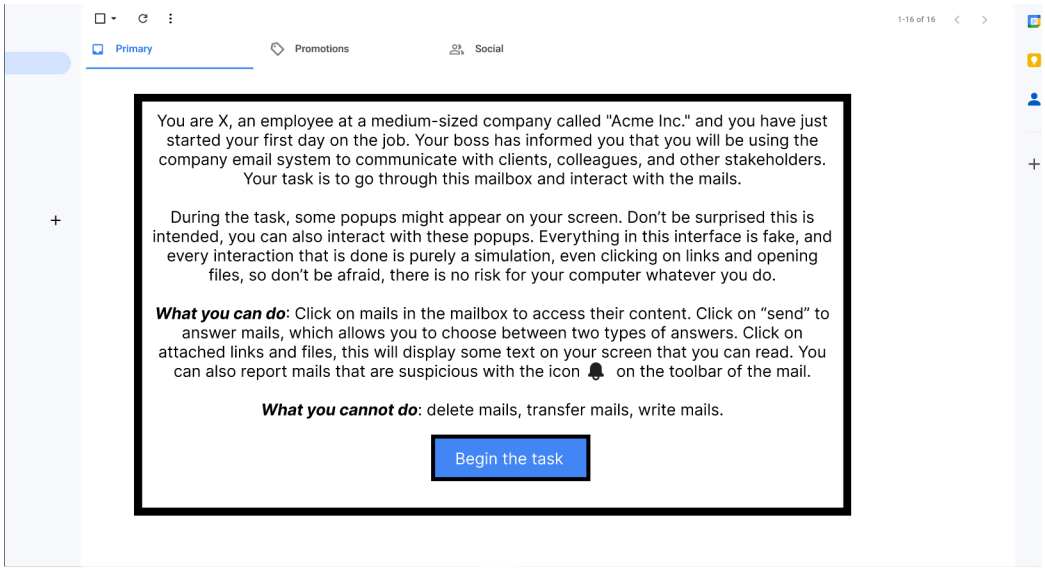

Figure 4: Scenario of background task

While the background task itself did not assess participants directly on their phishing-detection abilities, it did include one phishing email, as seen in 5. This initially had the purpose of measuring the number of participants who clicked on the phishing link. Unfortunately, Figma does not offer a built-in feature to track individual users visiting specific frames within a prototype. Therefore, the interactive elements served purely as interactive components without data collection capabilities.

Another such example was the "alert" buttons that I designed to enable participants to report suspicious emails. However, upon clicking the button, only a confirmation overlay was displayed, as seen in figure 6. The reported emails were not actually recorded. Nevertheless, the intention behind this feature was to promote participants' adoption of cautious behaviors. The "alert" button had the potential to serve as a valuable measurement tool for evaluating participants' tendencies to report emails, as well as the occurrence of false positives and false negatives. However, as mentioned previously, due to the limitations of Figma, this idea had to be abandoned and the functionality of tracking and recording reports was not implemented.

Finally, to enhance interactivity, participants were given the option to respond to the emails, as seen in figure 7. As text input was challenging to implement in Figma, I designed overlays with pre-filled answer options, allowing participants to explore different response possibilities. Due to time constraints, the background task was not extensively developed, as the main focus of the study was

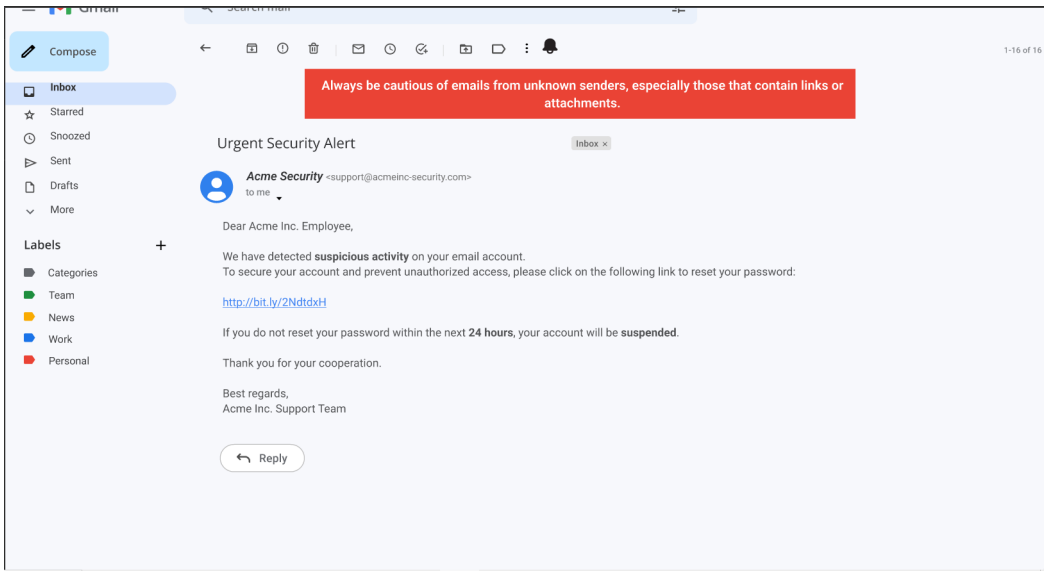

Figure 5: Phishing email in mockup

on the interventions. Additionally, designing the elements for the mockup required considerable time and effort. Nonetheless, learning how to correctly use Figma was also useful in order to create the phishing emails that would then serve for the questionnaires and training module.

3.3 Web app

To establish a logic chain between the questionnaires and the Gmail mockup, as well as to gather all the necessary data required for the analysis, the implementation of a web app was essential. Serving as the primary access point for participants, the web app was created and hosted on Glitch — a web-based platform that facilitates the development, collaboration, and hosting of web applications. Glitch offers an intuitive interface, real-time collaboration capabilities, built-in hosting services, and a strong emphasis on community engagement. Utilizing such a platform provides numerous advantages, notably its simplicity in deployment. However, one notable drawback is that it only supports server-side logic, which necessitates data collection through URL parameters and limits client interaction. The server-side scripting was done using JavaScript, employing various routes to handle different functionalities. The complete code for the *server.js* file that was used to run the server on Glitch can be found here.

Each participant was identified in the web app by their prolific ID. A SQLite database that contained a table called `Participants` was used to store each participant's data, including the prolific ID, the score and the category. Everytime a participant would access a certain route, the prolific ID was recovered from the URL parameter and looked-up in the table to retrieve the participant's information. An invalid prolific ID would display an error. The following functions were implemented to interact with the database:

- `initDB` to initialize the database and create the table `Participants`.
- `displayParticipants` to display the content of the `Participants` table.
- `addParticipant` to add a new participant to the table, where each participant is identified by their prolific id.
- `updateParticipant` to update a participant's field in the table given the prolific id and the field to update with the new value.

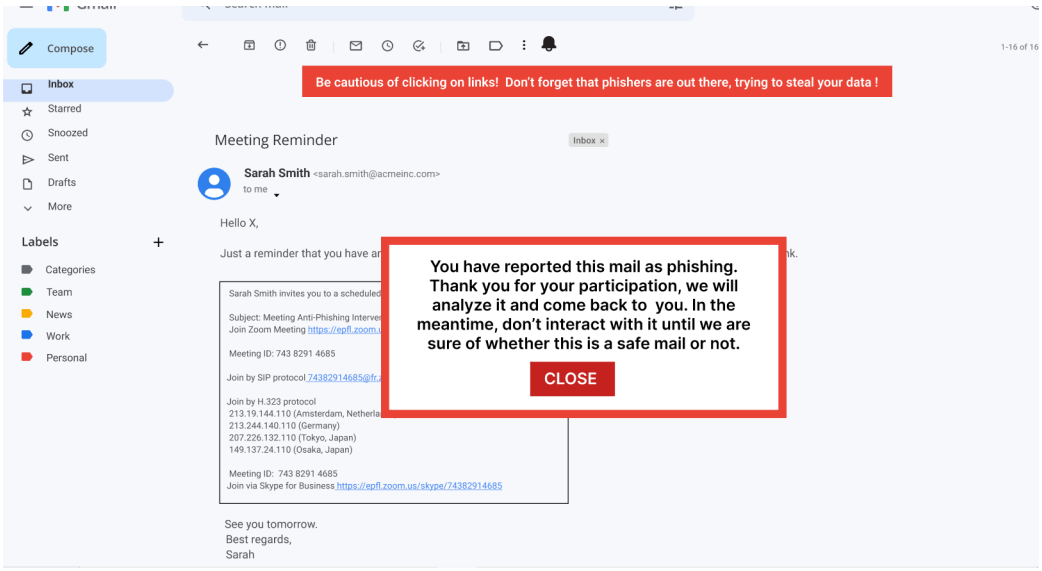

Figure 6: Report confirmation of phishing email

- `getElementByProlificId` that returns the row for the participant corresponding to the given prolific ID.
- `clearTable` to clear all entries of the `Participants` table.

Then, I implemented some helpers methods to compute the score and determine the user's category:

- `getPointsFromSelfReport` determined the bonus from the self-report.
- `computePracticalScore` computed the practical score based on the participant's answers to the phishing practice knowledge quizz. Only the phishing theory score was computed by Qualtrics and sent to the web app.
- `classifyUseFrequencyCategory` determined if the user had a rather high or a rather low frequency of use. This was used for scores in-between thresholds, as established in the scoring system.
- `classifyMailQuantityCategory` determined if the user had a rather high or a rather low email rate. This was also used for scores in-between thresholds, as established in the scoring system.
- `determineCategory` determined the category of the user, given the phishing theory score, the phishing practice score, the previous training experience vairable, the self report, the frequency of use and the email rate.

To implement the routes, I used `Fastify` which is a web framework for *Node.js* that focuses on providing high performance and low overhead for building web applications and APIs. It is designed to be highly efficient and scalable, making it suitable for handling heavy workloads and high traffic scenarios, which is perfect to handle multiple concurrent routes for a study. The following routes were used during the study:

- `/`: This route was the index. When receiving a HTTP GET request, the route redirected the user to the pre-intervention questionnaire on Qualtrics. As participants were starting the study from Prolific, this index was only used for testing and not in the final study.

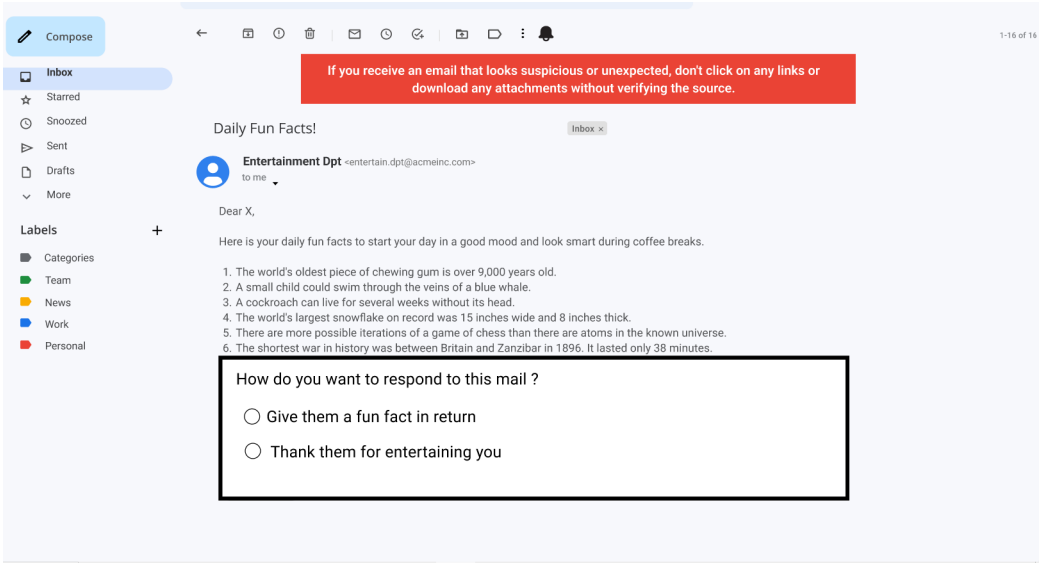

Figure 7: Answer box in Gmail mockup

- `/score`: This route was a POST-only route receiving requests from Qualtrics' web service. When the participants finished the pre-intervention questionnaire, their data were sent in the POST body to the `/score` route. The function was used to get all the data for a specific participant, identified by their prolific ID, determine their category and add the attributes to the Table with the `addParticipant` function.

- `/bgtask`: After transmitting the data to the `/score` route, Qualtrics promptly redirected the user to this page, incorporating the prolific ID as a URL parameter. Upon receiving the HTTP GET request, this route utilized the prolific ID to retrieve the participant's information from the Table. If the user did not match any valid participant in the database, meaning no participant was found with the provided id, they were redirected to the `/base_page` route, where a failure message was displayed. On the other hand, if a valid participant was identified based on the id, an HTML page was generated as a response, taking into account the participant's category. This HTML page included a frame that embedded the Figma mockup. The decision to use a frame for the mockup, instead of redirecting the participant to Figma directly, was made to enable the integration of interventions as pop-ups. These pop-ups were defined within the HTML code's script as overlays that appeared after a specified duration. The timing of the pop-ups varied depending on the participant's category. For instance, in Group 0, the education pop-up emerged after 2 minutes, while in Groups 0 and 1, the training pop-up appeared after 5 minutes. Group 2 did not have any pop-ups, as the reminders were already incorporated within the mockup itself. Additionally, a variable called `bgtask_after_training`, initially set to 0, was employed to determine whether the participant had already undergone training or not. The training module, which will be further presented in the `3.4` section, was displayed as a pop-up in the frame. When clicking the button on the pop-up, participants were redirected to Qualtrics. As Figma does not offer a means to track users across sessions, this variable was utilized to ensure accurate timing for task completion. Participants with `bgtask_after_training` set to 1 were considered to have already completed the training, and the entire HTML response from the `/bgtask` route allowed only 1 minute before displaying the final pop-up, indicating the task's conclusion. This approach was implemented to ensure that no significant timing discrepancy occurred between the groups that received the training intervention (which, based on prior timing analysis of a sample, took approximately 8 minutes) and Group 2, which had no such intervention. Group 2 was interacting with the background task for 10 minutes. The allocated timing for Group 2 was intentionally set to be shorter than the expected time for Group 0 and 1. This decision was made to prevent participants from becoming bored or disengaged after completing the background task. Once the designated time for the background task had elapsed, a pop-up appeared notifying participants that the study had concluded. They were then redirected to the post-intervention questionnaire

on Qualtrics.

- `/base_page`: this page was displayed for user that didn't have a valid prolific ID. This was done in order to avoid any unintended user to access the website and tamper the data.
- `/end_training`: This route was called with HTTP POST by Qualtrics after the training module to set the `bgtask_after_training` variable to 1 for a specific participant, using the `UpdateParticipant` method.
- `/end_study`: This route was intended to be called with HTTP POST at the end of the study, after the post-intervention questionnaire. The questionnaire from Qualtrics sent the new phishing theory score, and this route computed the new category after being exposed to the interventions, updating the participant fields in the table.

3.4 Interventions

The first intervention implemented in the study involved an educational module that was exclusively displayed to participants in Group 0. The purpose of this module was to convey fundamental concepts of phishing and common methods used in phishing attacks. The educational content was designed to be easily understandable, catering to individuals who may not be particularly tech-savvy or don't hold the patience to read a whole article about phishing. To maximize the amount of information conveyed in a short time frame and engage participants in a playful manner, a YouTube video titled "What is Phishing?" by Topic Simple, which can be viewed here, was considered an effective initial approach. This 2-minute and 10-second video effectively captured the core aspects of phishing and provided tips for detecting it, making it an ideal introduction to the topic. It comprehensively addressed various aspects of phishing, including its definition, the reasons behind its effectiveness, preferred channels used by attackers, common characteristics, ideal circumstances for successful phishing attempts, and recommended actions to take when encountering a suspicious email. Figure 8 shows the pop-up containing the Youtube video frame. Participants were prompted to watch the video as the overlay prevented any interaction with the Figma frame. Only after 2 minutes, a button was displayed on the pop-up, allowing participants to close the overlay, assuming they had sufficient time to watch the video.

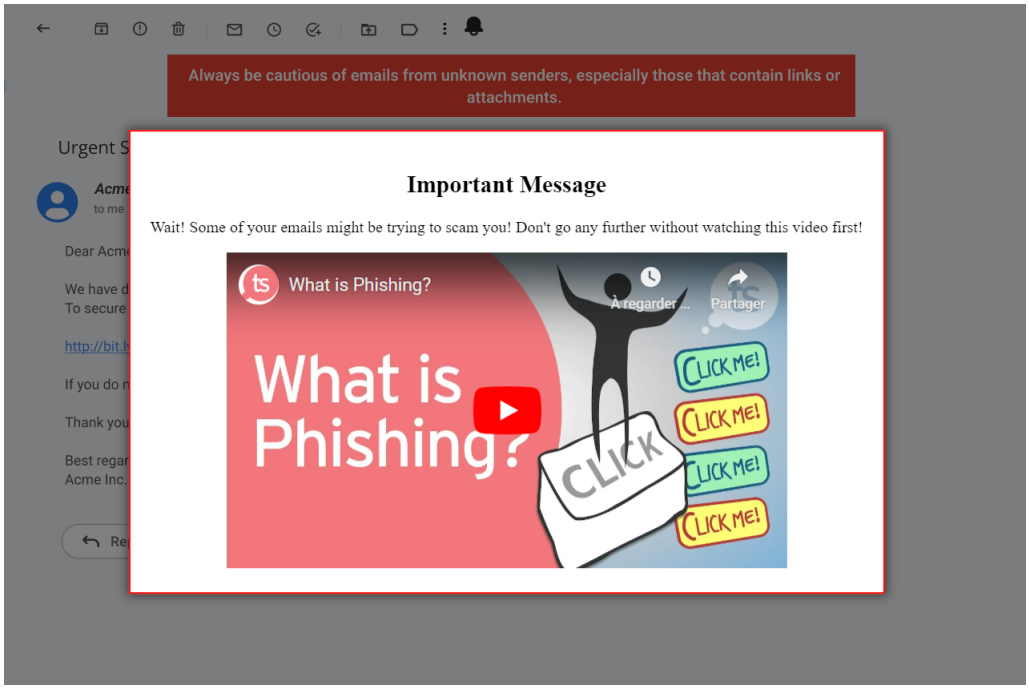

Figure 8: Education module

The second intervention consisted in a training module, which targeted participants in Group 0 and Group 1. The objective of this module was to train participants on techniques for recognizing phishing emails. To make the training interactive, a questionnaire was created using Qualtrics. Participants were presented with a series of emails, similar to the one shown in Figure 9, and were required to determine whether each email was legitimate or a phishing attempt. The training module aimed to familiarize participants with the detection techniques outlined in the section `2.1`, covering various aspects of phishing theory. Out of the eight emails provided, only one was legitimate, while the remaining emails were phishing attempts. It is important to note that this module was not designed to precisely replicate the distribution of legitimate and phishing emails found in real-world scenarios. Instead, its purpose was to train participants in improving their skills for detecting phishing attempts. The focus of the module was on enhancing participants' ability to recognize and differentiate between legitimate and fraudulent emails, rather than precisely mirroring the proportions observed in real-world situations. The emails covered the following aspects:

- `Suspicious link`: The first phishing email was meant to train the participants on their ability to analyze an URL and detect suspicious elements, such as an unknown or misspelled domain.
- `Sender address`: The second phishing email was meant to train the participants on their ability to analyze the sender's address by recognizing suspicious email domains for the alleged company.
- `Misspellings`: The third phishing email was meant to train participants on identifying misspellings. Misspellings often serve as strong indicators that the email may not originate from the claimed company.
- `Attached file extension`: The fourth phishing email was meant to train participants on identifying extensions for attached files that could indicate a malicious nature.
- `Reciprocity`: The fifth phishing email was meant to train participants on exercising caution when encountering content that appears too good to be true, as well as identifying specific elements that could be indicative of a phishing.
- `Scarcity`: The sixth phishing email was meant to train participants on exercising caution when encountering emails that contain alarming content, as well as identifying specific elements that could suggest a potential phish.
- `Authority`: The seventh phishing email was meant to train participants on exercising caution when encountering emails that claim to be from a powerful entity requiring something personal from them.
- `Legitimate email`: The final email was legitimate, in order to provide a concluding message about the importance of striking a balance between caution and not being overly suspicious. Its purpose was to emphasize the need for participants to develop effective tools for detecting phishing attempts rather than reacting impulsively or without proper judgment.

The emails used throughout the study were created using Figma and drew inspiration from real-life phishing scenarios. Both the training emails and explanations of the techniques utilized can be accessed here. The interactive nature of the module, which allowed participants to click on "Yes" or "No," was designed to encourage active engagement and promote analysis of the email content. This approach, resembling a game-like experience, was hypothesized to be more effective than presenting the participants with plain text, as it stimulated their critical thinking and decision-making processes. Upon completing the training module, participants were provided with a concise summary, aimed at reinforcing the key takeaways of the training. As explained in the section `3.3`, participants were redirected from the background task to the training module on Qualtrics after 5 minutes. After completion of the training, they would again be redirected to the background task for one last minute of interaction before being redirected to the post-intervention questionnaire.

The third intervention involved the implementation of reminders, which consisted of simple warning messages embedded directly within the gmail mockup, as depicted in 10. All groups were exposed to

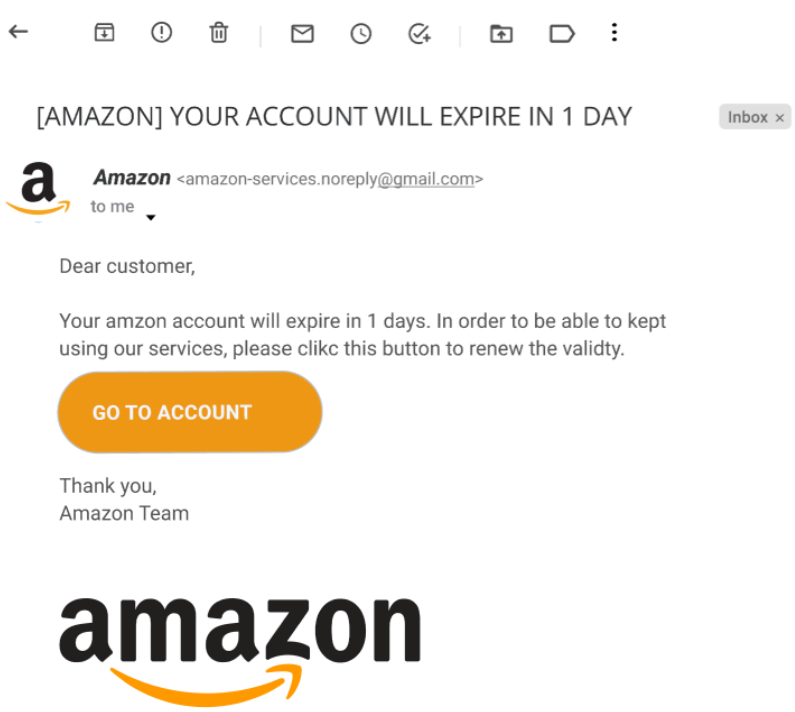

Figure 9: Example of phishing email used for the training module

these subtle interventions. The objective was to assess the effectiveness of reminders, particularly in Group 2. The education module was expected to be highly effective for Group 0 and the training module was expected to be beneficial for both Group 0 and Group 1. However, reminders were not expected to have a significant impact on these groups. The primary focus was on Group 2, as they supposedly already possess knowledge and training, and it was assumed that gentle reminders to exercise caution would generally help them increase their performance by raising their awareness. The reminders deliberately avoided providing specific information and remained broad in nature, such as advising participants to refrain from clicking on unknown links while not explaining why they shouldn't and without providing any tip on how they can be sure that a link is safe. This approach was primarily intended to maximize effectiveness in Group 2 specifically, as they were already supposed to know the techniques, while not providing any additional help to the other groups. The warning messages were present in each email of the mockup and were all different. Moreover, the subject of the reminders was randomized to ensure that they were not necessarily linked to the email's content. This approach aimed to prevent participants from being influenced in the background task by the presence of the reminders.

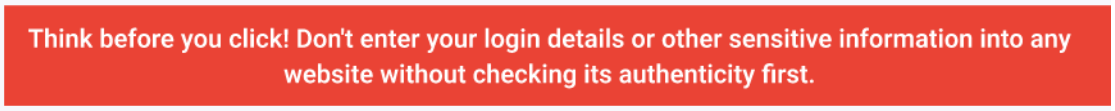

Figure 10: Reminder used during the background task

3.5 Technical challenges

The main technical challenges I encountered throughout the implementation were:

- Keeping track of participants: Since the implementation was limited to the server-side, tracking participants throughout the study posed a challenge due to multiple redirections. Initially, I attempted to utilize Fastify variables for this purpose. However, I encountered concurrency issues where simultaneous usage of the server by two participants would result in variable overwriting. To address this problem, I introduced a database. However, I encountered another challenge regarding participant identification. The only viable approach to track participants reliably was to utilize URL parameters consistently throughout the study. This ensured that each participant could be uniquely identified and their progress could be accurately monitored.

- Interactions with the Figma frame: Having the Figma prototype embedded as a frame raised concerns regarding potential issues, such as participants inadvertently clicking on unintended elements that could redirect them to the Figma hub and disrupt the study. To address this, I had to incorporate transparent overlays on sensitive elements within the frame. Despite being an effective solution, it could create confusion for participants that were not able to click on specific elements, which was not ideal. Additionally, I had initially intended to record certain interactions within the Figma frame; however, due to time constraints and limitations within Figma itself, such as the lack of click tracking features, I was unable to implement this recording functionality. Lastly, implementing the timer and pop-up mechanisms, while not excessively challenging, still required patience and effort to ensure their proper functionality within the study.

- Participants not being added to the database: After conducting the initial draft run with 10 participants, a significant technical issue arose whereby most of the participants didn't access the post-intervention questionnaire. This issue stemmed from a parsing bug within the `determineCategory` function, specifically related to the self_report variable. As a result, the function threw an exception for participants in Group 1 and Group 2, which resulted in these participants not being added to the database. Consequently, when the server ran the `/bgtask` route, it failed to retrieve information for participants that were not priorly added to the database. For these participants, the background task terminated after 1 minute. This was due to the route assuming that `bgtask_after_training` was set to 1, as it couldn't access any information from the database. To address this bug, I first corrected the parsing error and then implemented a solution to prevent similar issues in the future. I added a default case where, if the server failed to obtain data for a specific participant, default values would be assigned. For instance, participants would be automatically assigned to a random group with equiprobability, and `bgtask_after_training` would be set to 0 by default. Furthermore, another potential vulnerability was discovered prior to the full study when using special characters in the URL parameters. These characters were not properly encoded in the database, resulting in the prolific ID stored in the database differing from the correct ID. Although this issue was not significant, as prolific IDs are not expected to contain special characters, it is important to acknowledge and note its occurrence, which could have been solved with more time.

- Participants confusion: During the second run of the study, it was discovered that 20 participants did not complete the entire study, including the post-intervention questionnaire. Further investigation revealed that this issue was not due to a technical problem but rather stemmed from participant confusion regarding the completion time of the background task. To address this problem, a half-way pop-up was introduced after 5 minutes of interaction, as depicted in Figure 11. The purpose of this pop-up was to provide participants with guidance on the remaining duration of the task. Additionally, adjustments were made to the presentation text in the Figma mockup to clarify the expected completion time of the task. A sentence was added, stating: "The task will last approximately 10 minutes, so please take your time to thoroughly read and interact with all the elements, including searching for any hidden features. You will receive a final pop-up notification when the task is complete. Please refrain from refreshing the page or closing the tab until the task has ended." These modifications were implemented for a third run of the study, involving an additional 20 participants, with the aim of gathering more statistical power.

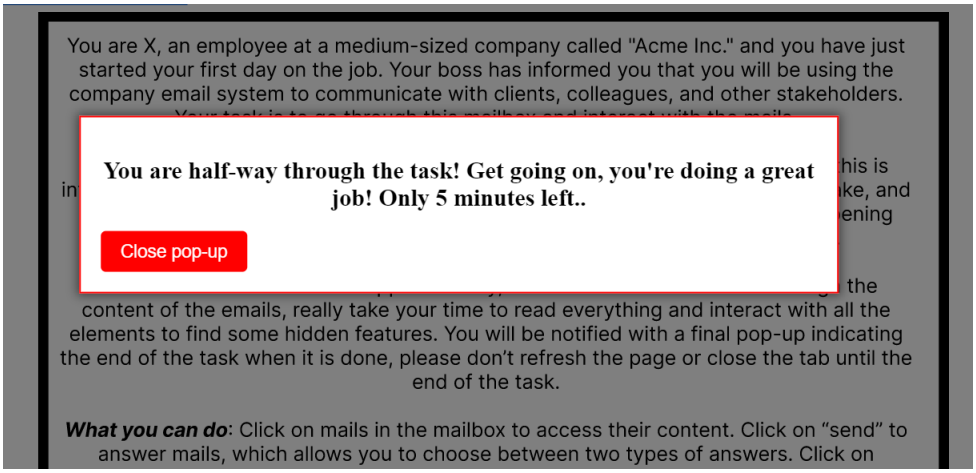

Figure 11: Halfway Pop-Up after 5 minutes of interaction

4 Results

The final phase of the study involved the examination and comparison of findings obtained from the questionnaires administered before and after the intervention. This analysis entailed performing a range of statistical tests and calculations. Details about the code for the analysis can be found here. In the subsequent section, I will present the outcomes derived from descriptive statistics, hypothesis testing and regression analysis. These results aim to ascertain notable enhancements in participants' scores and identify any correlations between human factors and the effectiveness of the interventions.

4.1 Descriptive statistics

Table 1 presents the proportions for each category before the participants had been exposed to the interventions and after having been exposed to the interventions. The interventions led to a decrease in the proportions of participants in Group 0 and Group 1, while there was an increase in the proportion of participants in Group 2.

Table 2 gives an overview of the descriptive statistics for each group before the interventions on the total score. This provides information on the average total score per group, as well as the standard deviation, minimum, maximum, 25th percentile and 75th percentile. Group 2 tends to have higher mean (which is expected as the classification is determined from the score), median and 25th/75th percentile scores, lower variability, and a wider range of pre-intervention total scores compared to Group 1 and Group 0. Group 0 generally has the lowest scores, higher variability, and a narrower range.

Table 3 shows the score differences by category, for the total score, the theory score and the practical score separately. The score differences are computed as the score post-intervention minus the score pre-intervention. Group 0 had the highest average score improvement, but also the highest variability, while Group 1 and Group 2 had lower average improvements with relatively lower variability.

Table 1: Category Proportion Pre-intervention and Post-intervention on 100 participants

Category	Pre-Intervention	Post-Intervention
Group 0	0.09	0.02
Group 1	0.35	0.10
Group 2	0.56	0.88

Table 2: Descriptive Statistics for Pre-intervention Total Score by Category

Category	Mean	SD	Median	Min	Max	25th Perc.	75th Perc.
Group 0	15.87	6.23	17.90	0.0	19.90	17.10	18.60
Group 1	34.77	3.10	35.32	28.50	39.15	32.37	37.67
Group 2	43.34	2.83	42.85	39.50	50.10	40.80	44.92

Table 3: Score Differences Statistics (Post-Intervention minus Pre-Intervention) by Category

	Total score diff			Theory score diff			Practical score diff		
Category	Mean	SD	C.I	Mean	SD	C.I	Mean	SD	C.I
Group 0	22.10	14.48	(10.98,33.23)	15.02	13.36	(4.75, 25.29)	0.89	2.67	(-1.16, 2.94)
Group 1	10.85	5.66	(9.34,12.37)	10.03	4.74	(8.76,11.30)	0.82	2.11	(0.25,1.39)
Group 2	6.69	4.40	(5.18,8.21)	6.98	3.87	(5.65,8.31)	-0.28	1.30	(-0.73,-0.16)

4.2 Hypothesis testing

To assess the effectiveness of the interventions, I conducted hypothesis tests to examine both the score differences and the changes in categories.

I performed a one-sample t-test to determine whether the score differences after the intervention were significantly different from zero for each category. The results indicated a significant improvement in scores for individuals in Group 0 (M = 22.10, SD = 14.48), t(8) = 4.581, p = 0.002, in Group 1 (M = 10.85, SD = 5.66),t(55) = 14.350, p = 0.000 and in Group 2 (M = 6.69, SD = 4.40), t(34) = 8.996, p = 0.000. These results provided evidence that the interventions had a positive impact on all category of individuals.

To assess whether there were significant changes in categories after the intervention, I conducted a chi-square test of independence. The results revealed a significant association between the pre-intervention and post-intervention categories ($\chi^2(2) = 21.169$, p = 0.000). This indicated that the interventions had a notable effect on category transitions. These findings support our hypothesis that the interventions would significantly improve scores and lead to changes in categories, i.e individuals from lower groups tends to transition to higher groups after being exposed to the interventions.

To assess the differences in scores between the groups, I conducted independent two-sample t-tests. The results are summarized below:

- The t-test results revealed a significant difference in scores between Group 0 and Group 1 (t(63) = 4.241, p = 0.000). This suggests that the scores of Group 0 are significantly different from those of Group 1. Based on the average total score being higher for Group 0, we can conclude that Group 0 had a more notable increase of score than Group 1 with the interventions.
- The t-test results revealed a significant difference in scores between Group 1 and Group 2 (t(89) = 3.701, p = 0.000). This suggests that the scores of Group 1 are significantly different from those of Group 2. Based on the average total score being higher for Group 1, we can conclude that Group 1 had a more notable increase of score than Group 2 with the interventions.
- The t-test results indicated no significant difference in scores between Group 0 and Group 2 (t(42) = 5.530, p = 0.000). This suggests that the scores of Group 0 are significantly different from those of Group 2. Based on the average total score being higher for Group 0, we can conclude that Group 0 had a more notable increase of score than Group 2 with the interventions.

Therefore, the interventions exhibited significant efficacy in improving scores and categories overall. Additionally, the findings suggest that the interventions were more effective for participants in lower groups.

4.3 Regression analysis on demographic factors

4.3.1 All groups

I explored the potential correlation between demographic factors (namely age group, education level, IT security background and occupation) and the total score difference before and after the interventions by conducting an OLS regression on all groups. Note that only the groups before the interventions were evaluated. The R-squared value for the OLS regression was 0.191, indicating that the model explained approximately 19.1% of the variance in the total score difference. The adjusted R-squared value, which takes into account the number of predictors, was -0.013. This suggested that the selected predictors may not have been effectively explaining the variation in the total score difference. Based on results in table 4,I observed that all p-values were higher than 0.05, meaning that no variable was significant for the total score difference. This means that demographic factors in general don't have any impact on the efficacy of the interventions.

To evaluate whether some demographic factors could be correlated to the efficacy of a specific intervention, I then conducted OLS regressions on each Group separately.

Table 4: OLS Regression Results on all groups

Variable	**Coefficient**	**Std. Error**	**p-value**
const	2.0000	7.720	0.796
age:18-25	0.6188	3.101	0.842
age:26-35	1.5994	2.675	0.552
age:36-45	1.6076	2.710	0.555
age:46-55	1.8133	2.702	0.504
age:56-65	-2.0506	3.413	0.550
age:66-75	3.2589	6.193	0.600
education:Associate degree	-0.5936	4.773	0.901
education:Other	2.8968	4.290	0.501
education:PhD or similar	-1.1393	5.422	0.834
education:Secondary school diploma	2.8722	2.961	0.335
education:University degree	2.8112	2.692	0.300
IT background:No	2.6642	3.139	0.399
IT background:Yes (IT specialist)	-5.0130	5.609	0.374
IT background:Yes (other IT security)	2.6123	4.017	0.517
IT background:Yes (studies in CS)	6.5840	3.832	0.090
occupation:Civil Service	7.2501	3.876	0.065
occupation:Employee	1.5617	2.138	0.467
occupation:Other	0.3112	4.599	0.946
occupation:Retired	3.9930	4.838	0.412
occupation:Self-employed	2.5821	2.776	0.355
occupation:Training/ Apprenticeship	-3.0916	5.469	0.573
occupation:Unemployed	-7.9888	4.835	0.102
occupation:University Student	2.2298	3.270	0.497

4.3.2 Group 0

The OLS regression results on Group 0 indicated a perfect fit to the data with an R-squared value of 1.000, suggesting that the model explained all the variability in the total score difference. The adjusted R-squared of 0.998 further supported the strong fit of the model. The F-statistic of 748.3 with a p-value of 0.0281 indicated that the overall model was statistically significant, suggesting that at least one of the predictor variables was related to the total score difference. Based on results in table 5, I observed the following for the demographic factors:

Table 5: OLS Regression Results on Group 0

Variable	Coefficient	Std. Error	p-value
const	8.4220	0.145	0.011
age:26-35	-3.6244	0.185	0.033
age:36-45	1.0846	0.297	0.170
age:46-55	10.9619	0.314	0.018
education:Associate degree	-5.4699	0.350	0.041
education:Secondary school diploma	5.6028	0.349	0.040
education:University degree	8.2891	0.210	0.016
IT background:No	3.7271	0.340	0.058
IT background:Yes (IT specialist)	-13.7729	0.531	0.025
IT background:Yes (other IT security)	6.5544	0.312	0.030
IT background:Yes (studies in CS)	11.9135	0.185	0.010
occupation:Civil Service	11.9135	0.185	0.010
occupation:Employee	10.6362	0.548	0.033
occupation:Retired	-11.8138	0.598	0.032
occupation:Self-employed	-2.3138	0.324	0.088

- Age: Being part of the 26-35 group had a significant negative impact on the total score difference, while being in the 46-55 group had a significant positive impact.
- Education: Having a Secondary school diploma or an University degree had a significant positive impact on the total score difference, whereas having an Associate degree had a significant negative impact on the total score difference.
- IT Background: Answering "Yes, other IT security" or "Yes, studies in Computer Science" had a positive and significant impact on the total score difference. Being an IT specialist surprisingly had a negative impact on the total score difference for Group 0. We can't say with confidence whether not having an IT background impacted or not the total score difference.
- Occupation: Being at Civil service or an Employee has a significant positive impact on the total score difference. Being Retired has a significant negative impact on the total score difference. We can't say with confidence whether being Self employed has an impact or not.

4.3.3 Group 1

For Group 1, the regression model had an R-squared value of 0.402, indicating that approximately 40.2% of the variability in the total score difference could be explained by the independent variables in the model (demographic factors). However, the adjusted R-squared value was 0.086, which suggested that when considering the number of predictors in the model, the explanatory power was lower. Furthermore, as we can see in table 6, no variable had a statistically significant coefficient. Group 1 being the largest group, this also accentuates the idea that demographic factors are for the most population not determinant of the performance.

4.3.4 Group 2

For Group 2, the R-squared value was 0.566, which means that approximately 56.6% of the variance in the total score difference could be explained by demographic factors. The adjusted R-squared value was 0.223 in this case, indicating that when considering the number of predictors and degrees of freedom, approximately 22.3% of the variance in the total score difference was explained by the demographic factors. Based on results in table 7, I observed the following:

- Education: education level "University" had a coefficient of 5.3813 and a low p-value (0.006), suggesting a statistically significant positive relationship with the total score difference

Table 6: OLS Regression Results on Group 1

Variable	Coefficient	Std. Error	p-value
const	2.0000	5.411	0.714
age:18-25	3.0142	3.115	0.340
age:26-35	2.8483	2.576	0.276
age:36-45	2.8763	2.449	0.248
age:46-55	-1.4378	2.431	0.558
age:56-65	-3.8164	4.030	0.350
age:66-75	3.5404	5.642	0.534
education:Associate	-2.7690	7.099	0.699
education:Other	2.2642	3.305	0.498
education:PhD	-0.6967	4.325	0.873
education:Secondary	3.7334	2.652	0.168
education:University	4.4931	2.316	0.060
IT background:No	2.1604	3.103	0.491
IT background:Yes	1.2954	5.024	0.798
IT background:Yes CS	3.5691	4.520	0.435
occupation:Civil	0.3841	4.459	0.932
occupation:Employee	1.6196	2.482	0.518
occupation:Other	4.4658	5.261	0.402
occupation:Retired	3.6226	5.533	0.517
occupation:Self	2.2491	2.874	0.439
occupation:Training	3.5756	7.078	0.617
occupation:Unemployed	-9.8677	5.549	0.084
occupation:University	0.9759	3.215	0.763

- IT background: Answering "Yes, IT background" had a coefficient of 8.3615 and a p-value of 0.041, indicating a statistically significant positive relationship.
- Occupation: "University" had a coefficient of 6.5257 and a p-value of 0.014, suggesting a statistically significant positive relationship.

4.4 Regression analysis on personality traits

4.4.1 Relationship between personality traits and total score

An OLS regression was conducted to examine the connection between personality traits and the total score obtained from the pre-intervention questionnaire. The personality traits were determined based on the responses to the personality questions in the post-intervention questionnaire. The traits assessed included the following: introverted, trusting, indolent, serene, not artistic, sociable, intolerant, conscientious, anxious and creative. The R-squared value was 0.363, indicating that approximately 36.3% of the variation in the total score could be explained by the personality traits. However, the adjusted R-squared value was -0.069, suggesting that the personality traits may not effectively account for the variation in the total score and that the model could be overfitting the data.

Analyzing the significant coefficients from the OLS results, it was found that strongly agreeing with the question on sociability had a significant negative impact (-11.0730) on the total score. This implies that individuals who strongly identified as sociable tended to perform worse on the pre-intervention questionnaire, hence were most susceptible to phishing.

Table 7: OLS Regression Results on Group 2

Variable	Coefficient	Std. Error	p-value
const	2.2563	1.220	0.080
age:18-25	-0.2619	1.658	0.876
age:26-35	-1.1434	1.427	0.433
age:36-45	2.7904	1.508	0.080
age:46-55	0.2719	1.711	0.875
age:56-65	0.5993	2.079	0.776
education:Associate	-1.5353	3.737	0.686
education:Secondary	-1.5897	2.222	0.483
education:University	5.3813	1.734	0.006
IT background:No	-2.9189	1.527	0.071
IT background:Yes (IT specialist)	8.3615	3.819	0.041
IT background:Yes (other IT security)	-2.9119	2.350	0.230
IT background:Yes (studies in CS)	-0.2744	2.179	0.901
occupation:Civil	1.6948	2.983	0.577
occupation:Employee	2.1153	1.364	0.137
occupation:Other	-1.3091	3.763	0.732
occupation:Self	-4.3382	4.214	0.316
occupation:Training	0.3247	3.824	0.933
occupation:Unemployed	-2.7568	3.918	0.490
occupation:University	6.5257	2.398	0.014

4.4.2 Relationship between personality traits and new total score

Another OLS regression was performed to examine the association between personality traits and the new total score, which represents the score obtained from the post-intervention questionnaire. The R-squared value was 0.462, indicating that approximately 46.2% of the variance in the new total score could be explained by the personality traits. However, the adjusted R-squared value was 0.098, suggesting that the personality traits may not effectively explain the variation in the new total score.

The results revealed that disagreeing slightly with the question on sociability had a significant positive impact on the new total score. This implies that individuals who were not very sociable were more likely to perform better on the post-intervention questionnaire. This also corrobates the idea that sociability may have an impact on the ability to recognize phishing.

4.4.3 Relationship between personality traits and total score difference

Further analysis was conducted to explore the potential relationship between personality traits and the difference in total scores by running an OLS regression. The R-squared value was 0.300, indicating that the personality traits explained 30% of the variance in the total score difference, suggesting a moderate fit of the model. However, the adjusted R-squared value was -0.175, indicating a possibility of overfitting the data.

Most of the personality traits did not show statistically significant coefficients, suggesting that they were not significantly related to the total score difference. Among the significant coefficients, strongly agreeing with trust had a coefficient of 8.1333, indicating that individuals who strongly agreed with this personality trait, on average, had a higher total score difference compared to those who did not. Conversely, strongly disagreeing with trust had a coefficient of -13.2580, indicating that individuals who strongly disagreed with this personality trait, on average, had a lower total score difference compared to those who did not. The findings suggest that trustful individuals may benefit more from the interventions, as they exhibited larger improvements in their scores after the intervention. This implies that the efficacy of the interventions can be more impactful for individuals who have a higher level of trust.

4.4.4 Relationship between personality traits and total score difference on separate groups

I then looked at the relationship between personality traits and total score difference for each group separately.

For Group 2, the R-squared value was 1.000, which means that the model explained 100% of the variance in the total score difference. This indicated a perfect fit, which could suggest potential issues with the model or data. All the coefficients had p-values of 0.000, which suggests that all personality traits have a statistically significant impact on the total score difference. Moreoever, the confidence intervals for each coefficient are very narrow, suggesting high precision in the estimates. However, it is important to note that the results showing an R-squared of 1.000 and p-values of 0.000 may indicate potential issues with the regression model, such as overfitting or perfect multicollinearity. These issues could arise due to the small sample size of Group 2 and may invalidate the findings.

For Group 1, the R-squared value of 0.732 indicated that the model explained approximately 73.2% of the variance in the total score difference. This suggested a moderately strong relationship between the personality traits and the total score difference. Agreeing strongly with the personality trait "introverted" had a coefficient of 7.0594. On the other hand, agreeing strongly with the personality "indolent" had a coefficient of 14.8113, suggesting an even stronger positive association with the total score difference.

5 Discussion

Effectiveness of interventions: The interventions were found to be generally effective with statistical confidence, significantly improving the performance of all groups. Notably, Group 0, exposed to all three interventions (education, training and reminders), exhibited the most substantial improvement. Group 1 had a smaller improvement but still significant, while Group 2, exposed only to reminders, showed the smallest but a still significant improvement. These results support the effectiveness of the proposed methodology and indicate successful interventions in enhancing participants' ability to detect phishing. Further investigation could be conducted to assess the outcomes of combining interventions for all three groups and determine if a particular intervention type could be universally effective. For example, combining education and training for Group 1 might lead to highly effective results. Such insights would contribute to refining and optimizing intervention strategies to maximize their effectiveness across different groups.

Demographic factors: The analysis of demographic factors revealed that, overall, these factors may not have a significant impact on the efficacy of the interventions, as indicated by the OLS regression conducted on all groups. This suggests that personalized interventions based on demographic characteristics may not yield fruitful results. Instead, the degree of knowledge appears to be the primary influencing factor, as hypothesized in Section `2.1`. Moreover, it is worth noting that the selected predictors in the model explained only a small proportion of the variance in the total score difference, indicating the presence of other unaccounted factors influencing the outcome. However, within individual groups, certain demographic factors demonstrated influence.
In Group 0, age, education level, and occupation exhibited significant relationships with the total score difference. Specifically, individuals in the age group 46-55 experienced a positive impact from the interventions, while those in the age group 26-35 experienced a negative impact. These findings suggest that the combination of interventions may not be well-suited for the 46-55 age group, and alternative tailored interventions could be explored. Furthermore, the interventions had a positive impact on individuals with an IT background, while IT specialists experienced negative effects. This unexpected finding suggests a possible mismatch between the intervention content and the existing knowledge and skills of IT specialists. The classification of IT specialists in the lowest group indicates that their specialization in IT may not be directly relevant to their ability to detect phishing. In contrast, individuals with a background in IT demonstrated flexibility in effectively integrating the

intervention content. This can be attributed to their broader knowledge and less specialized expertise, which may enable them to adapt and learn new areas within IT, such as phishing.
In Group 1, representing the largest portion of the sample, none of the demographic factors showed statistical significance in predicting the total score difference. This implies that demographic factors may not reliably predict the performance within this group. Considering the participants' level of knowledge as a significant factor would be more appropriate, given the complexity and potential conflicts of various factors within this group.
In Group 2, being in university exhibited a statistically significant positive relationship with the total score difference. This finding suggests that reminders served as effective prompts or cues for individuals with a deeper understanding of the subject matter, enabling them to recall and apply their existing knowledge more effectively. Factors such as comprehensive and in-depth learning experiences typically associated with university education, as well as the structured learning environment and self-directed learning approach of university students, may have contributed to their responsiveness to reminders and integration of the intervention content.
Based on the findings, it can be concluded that tailored interventions could be personalized based certain on demographic factors such as age, education, and IT background. However, it is generally recommended to primarily focus on personalization according to knowledge classification. The effectiveness of the interventions was influenced by demographic factors primarily for the most extreme knowledge groups (lowest and highest knowledge). Therefore, for sub-classifications of individuals, it may be beneficial to implement a two-layered approach to personalization. The first layer would involve classifying individuals based on their knowledge level, and the second layer would incorporate demographic factors. This approach would allow for a more targeted and effective implementation of interventions for specific subgroups.

Personality traits: The analysis of personality traits revealed meaningful insights into their impact on vulnerability to phishing attacks and the effectiveness of interventions. The findings suggest that specific traits can significantly influence individuals' susceptibility to phishing attempts and the outcomes of intervention strategies.
For instance, in group 1, the trait of introversion was found to be associated with a higher efficacy of interventions. This implies that individuals who identify as introverted may respond more positively to interventions such as training and reminders, making them more receptive to improving their knowledge and skills related to phishing. Similarly, the trait of indolence showed an even stronger efficacy for the interventions in Group 1.
For all groups, sociability emerged as another trait with an impact on susceptibility to phishing. The results indicate that individuals who are less sociable tend to perform better on questionnaires, indicating a lower susceptibility to falling for phishing scams. On the contrary, highly sociable individuals exhibited a negative impact on their pre-intervention scores, suggesting a higher vulnerability to phishing attempts. These findings reinforce the notion that sociability should be taken into account when designing personalized interventions to mitigate susceptibility to phishing attacks.
Regarding the effectiveness of interventions, the results highlight the role of trust as a significant personality trait for all groups. Trusting individuals showed greater improvements in their scores after interventions. Conversely, individuals that tend to be less trusting showed smaller improvements. This suggests that individuals who possess a higher level of trust may derive more benefits from interventions, as they are more likely to trust the information and guidance provided during the intervention process. It is important to note that while these correlations between personality traits, susceptibility to phishing, and intervention efficacy are observed, they are not universally significant for all cases. Overall, personality traits are just one aspect of the complex factors that influence vulnerability to phishing, and a comprehensive approach is necessary to address this issue effectively. Considering personality traits alongside other factors, such as demographic characteristics, can inform the development of targeted and personalized interventions to maximize their effectiveness.

6 Limitations

The interpretation of the results should be approached with caution due to several limitations. Firstly, in the general case, both demographic factors and personality traits were found to be statistically insignificant. This could be attributed to potential overfitting of the models or the presence of unexplained variance, indicating that these factors alone may not fully capture the complexity of susceptibility to phishing. It is crucial to consider various other factors that could contribute to individuals' vulnerability. Furthermore, the sample sizes used in the study were relatively small, which may have limited the statistical power and generalizability of the findings. With 10 different personality traits, each with a degree of 4, the inclusion of 40 independent variables in the regression analysis with only 100 participants might have compromised the accuracy of the results. Due to these limitations, regressions were conducted in parallel for personality traits and demographic factors rather than testing all factors together, which is not ideal for drawing definitive conclusions. Conducting further analyses, such as diagnostic checks for assumptions like normality and homoscedasticity, could provide valuable insights into the validity of the regression models.

It is also important to note that the participants were not fully controlled, leading to certain drawbacks in the study. Firstly, the distribution of participants among the groups was highly imbalanced, with the majority falling into Group 1, while Group 0 and Group 2 had fewer participants. Despite efforts to adjust the thresholds, achieving a balanced 33% distribution across the groups proved challenging. This imbalance arose from the fact that the classification aimed to mimic a real-world scenario, where Group 0 represented individuals lacking phishing knowledge. However, in the context of a study conducted through Prolific, where most participants had prior exposure to similar studies, the number of participants falling into Group 0 was significantly reduced. Similarly, Group 2, representing individuals with the highest level of knowledge, was difficult to populate with individuals lacking substantial knowledge. Adjusting the thresholds extensively could have resulted in a more balanced distribution among the groups, but it would have compromised the targeting of specific types of individuals, which was a key objective of the interventions. Additionally, some groups lacked a complete range of factors. For example, Group 0 included only specific age groups (26-35, 36-45, and 46-55), limiting the ability to assess the impact of interventions on other age groups. This suggests the need for complementary studies that investigate interventions separately for each group.

Additionally, the scoring system and classification methodology employed in the study may have introduced biases and deviated from real-world scenarios. Despite efforts to create a layered classification based on previous studies, there were evident flaws that require correction in future research. For instance, in the practical knowledge quiz, participants had only two options ("phishing" or "not phishing"), and the tendency to answer affirmatively due to the Hawthorne effect might have influenced the computation of the practical knowledge scores. Negative points were not considered, although including them could have balanced the expected score to zero for random answers. The decision to exclude negative points aimed to simplify the scoring process and avoid introducing additional complexity to the scoring system, but a simple scoring system for classification might not truly encapsulates how phishing susceptibility should be assessed.

Finally, the study could have benefited from a more realistic context, such as considering industry settings. Certain industries may be more susceptible to spear phishing attacks, which are distinct from classic phishing attempts. The efficacy of interventions may vary depending on the type of phishing targeted. If spear phishing poses a significant concern in specific industries, a more tailored and industry-specific intervention approach may be necessary. Addressing spear phishing as a distinct and critical issue is crucial, considering that phishing serves as a primary entry point for dangerous attacks and malware targeting organizations.

7 Conclusion

This study aimed to evaluate the effectiveness of personalized interventions based on knowledge classification and explore the correlation between various human factors and the efficacy of these interventions. The literature review highlighted knowledge as the primary influencing factor for susceptibility to phishing and a reliable indicator of an individual's category. The methodology section provided insights into the classification process and study implementation. The results demonstrated the overall high efficacy of the interventions across all groups, with lower groups demonstrating more improvement. Specifically, the educational module proved to be particularly effective for individuals with low knowledge, indicating its efficiency in improving their awareness. Training and reminders were effective for individuals with moderate knowledge, while reminders alone were also effective for those with high knowledge. However, further investigation is needed to determine whether different combinations of interventions across groups yield varying degrees of efficacy to identify the optimal intervention strategy. Although the correlation between demographic factors and intervention efficacy was not found to be significant overall, certain factors were relevant for the lowest and highest groups. Age, for example, influenced the impact of interventions on individuals with low knowledge, suggesting the need to tailor educational modules to different age groups. Additionally, education and IT background were worth considering when personalizing interventions. Designing interventions with a multi-layered classification approach is recommended, starting with knowledge classification and then incorporating demographic factors for more detailed and specific personalization. Future studies should investigate the specific effects of interventions on demographic factors to determine the most effective types of interventions. Similarly, while personality traits demonstrated limited significance, trust emerged as an important trait to consider when assessing intervention efficacy, and sociability remained a significant factor influencing susceptibility to phishing. Further research should delve deeper into the specific effects of interventions on demographic factors and explore additional personality traits to enhance the understanding and effectiveness of personalized interventions in combating phishing threats.

References

Data analysis code. URL https://github.com/Walfar/anti-phishing-intervention-tool/blob/main/anti-phishing-tool-data-analysis.ipynb.

Youtube video for education module. URL https://www.youtube.com/watch?v=WG8V1_Sj5g0.

Figma prototype for the gmail mockup. URL https://www.figma.com/proto/iQ1dxn6XosIPUm2xJglJxD/Anti-phishing-Gmail-Mockup?type=design&node-id=0-2&scaling=min-zoom&page-id=0%3A1&starting-point-node-id=912%3A881.

Github with study materials. URL https://github.com/Walfar/anti-phishing-intervention-tool.

Figma template used for the gmail mockup. URL https://www.figma.com/community/file/984693678888824118.

10 most common signs of a phishing email, a. URL https://cofense.com/knowledge-center/signs-of-a-phishing-email/.

How to spot a phishing email: With examples, b. URL https://www.itgovernance.co.uk/blog/5-ways-to-detect-a-phishing-email.

How to identify a phishing email, c. URL https://www.cloudflare.com/learning/email-security/how-to-identify-a-phishing-email/.

Qualtrics post-intervention questionnaire. URL https://github.com/Walfar/anti-phishing-intervention-tool/blob/main/post-intervention_questionnaire.pdf.

Qualtrics pre-intervention questionnaire. URL https://github.com/Walfar/anti-phishing-intervention-tool/blob/main/pre-intervention_questionnaire.pdf.

Javascript code for the web app server. URL https://github.com/Walfar/anti-phishing-intervention-tool/blob/main/server.js.

Training module content. URL https://github.com/Walfar/anti-phishing-intervention-tool/blob/main/training_module.pdf.

Rahmad Abdillah, Zarina Shukur, Masnizah Mohd, and Ts. Mohd Zamri Murah. Phishing classification techniques: A systematic literature review. *IEEE Access*, 10:41574–41591, 2022. doi: 10.1109/ACCESS.2022.3166474.

A. Alnajim and Malcolm Munro. Effects of technical abilities and phishing knowledge on phishing websites detection. pages 120–125, 01 2009.

Gitanjali Baral and Nalin Asanka Gamagedara Arachchilage. Building confidence not to be phished through a gamified approach: Conceptualising user's self-efficacy in phishing threat avoidance behaviour. pages 102–110, 2019. doi: 10.1109/CCC.2019.000-1.

Casey Inez Canfield and Baruch Fischhoff. Setting priorities in behavioral interventions: An application to reducing phishing risk. *Risk Analysis*, 38(4):826–838, 2018a. doi: https://doi.org/10.1111/risa.12917. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.12917.

Casey Inez Canfield and Baruch Fischhoff. Setting priorities in behavioral interventions: An application to reducing phishing risk. *Risk Analysis*, 38(4):826–838, 2018b. doi: https://doi.org/10.1111/risa.12917. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.12917.

Jin-Hee Cho, Hasan Cam, and Alessandro Oltramari. Effect of personality traits on trust and risk to phishing vulnerability: Modeling and analysis. pages 7–13, 2016. doi: 10.1109/COGSIMA.2016.7497779.

Nippert-Eng C. Das, S. and L.J Camp. Evaluating user susceptibility to phishing attacks. pages 1–18, 2022. URL https://doi.org/10.1108/ICS-12-2020-0204.

Giuseppe Desolda, Lauren S. Ferro, Andrea Marrella, Tiziana Catarci, and Maria Francesca Costabile. Human factors in phishing attacks: A systematic literature review. *ACM Comput. Surv.*, 54(8), oct 2021. ISSN 0360-0300. doi: 10.1145/3469886. URL https://doi.org/10.1145/3469886.

Zuochao Dou, Issa Khalil, Abdallah Khreishah, Ala Al-Fuqaha, and Mohsen Guizani. Systematization of knowledge (sok): A systematic review of software-based web phishing detection. *IEEE Communications Surveys Tutorials*, 19(4):2797–2819, 2017. doi: 10.1109/COMST.2017.2752087.

Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. SoK: Still plenty of phish in the sea — a taxonomy of User-Oriented phishing interventions and avenues for future research. pages 339–358, August 2021. URL https://www.usenix.org/conference/soups2021/presentation/franz.

Yan Ge, Li Lu, Xinyue Cui, Zhe Chen, and Weina Qu. How personal characteristics impact phishing susceptibility: The mediating role of mail processing. *Applied Ergonomics*, 97:103526, 2021. ISSN 0003-6870. doi: https://doi.org/10.1016/j.apergo.2021.103526. URL https://www.sciencedirect.com/science/article/pii/S0003687021001733.

Matthew D Grilli, Katelyn S McVeigh, Ziad M Hakim, Aubrey A Wank, Sarah J Getz, Bonnie E Levin, Natalie C Ebner, and Robert C Wilson. Is This Phishing? Older Age Is Associated With Greater Difficulty Discriminating Between Safe and Malicious Emails. *The Journals of Gerontology: Series B*, 76(9):1711–1715, 12 2020. ISSN 1079-5014. doi: 10.1093/geronb/gbaa228. URL https://doi.org/10.1093/geronb/gbaa228.

Tzipora Halevi, James Lewis, and Nasir Memon. A pilot study of cyber security and privacy related behavior and personality traits. page 737–744, 2013. doi: 10.1145/2487788.2488034. URL https://doi.org/10.1145/2487788.2488034.

Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of phish: A real-world evaluation of anti-phishing training. 2009. doi: 10.1145/1572532.1572536. URL https://doi.org/10.1145/1572532.1572536.

Lina; Liu, Zhihui; Zhou and Dongsong Zhang. Effects of demographic factors on phishing victimization in the workplace. 2020. URL https://aisel.aisnet.org/pacis2020/75.

Pablo López-Aguilar and Agusti Solanas. Human susceptibility to phishing attacks based on personality traits: The role of neuroticism. pages 1363–1368, 2021. doi: 10.1109/COMPSAC51774.2021.00192.

Dawn M. Sarno and Mark B. Neider. So many phish, so little time: Exploring email task factors and phishing susceptibility. *Human Factors*, 64(8):1379–1403, 2022. doi: 10.1177/0018720821999174. URL https://doi.org/10.1177/0018720821999174. PMID: 33835881.

Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. page 373–382, 2010. doi: 10.1145/1753326.1753383. URL https://doi.org/10.1145/1753326.1753383.

Alex Sumner and Xiaohong Yuan. Mitigating phishing attacks: An overview. page 72–77, 2019. doi: 10.1145/3299815.3314437. URL https://doi.org/10.1145/3299815.3314437.

McKenna K. Tornblad, Keith S. Jones, Akbar Siami Namin, and Jinwoo Choi. Characteristics that predict phishing susceptibility: A review. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 65(1):938–942, 2021. doi: 10.1177/1071181321651330. URL https://doi.org/10.1177/1071181321651330.

Yuan; Wang, Jingguo; Li and H. Raghav Rao. Overconfidence in phishing email detection. 2016. doi: 10.17705/1jais.00442. URL https://aisel.aisnet.org/jais/vol17/iss11/1.