PickMail: A Serious Game for Email Phishing Awareness Training

Gokul Jayakrishnan
TCS Research,
Tata Consultancy Services Ltd.
gokul.cj@tcs.com

Vijayanand Banahatti
TCS Research,
Tata Consultancy Services Ltd.
vijayanand.banahatti@tcs.com

Sachin Lodha
TCS Research,
Tata Consultancy Services Ltd.
sachin.lodha@tcs.com

***Abstract*-** Phishing threats are on the rise, especially through Business Email Compromise (BEC). Despite having several tools for phishing email detection, the attacks are becoming smarter and personal, targeting individuals to gain access to personal and organizational information. Game-based cybersecurity training methods are found to have positive results in educating users. Along this line, we introduce *PickMail*, an anti-phishing awareness game that simulates typical real-life email scenarios to train an organization's employees. In *PickMail*, we train participants to judge the legitimacy of an email by inspecting its various parts, such as the sender's email domain, hyperlinks, attachments, and forms. The game also records participants' decision-making steps that lead to their final judgment. Our study with 478 participants shows how the serious game-based training helped the participants make better judgments on emails, with the correctness in identifying email legitimacy reaching 92.62%. The study also provided us with insights that could help develop better training methods and user interfaces.

I. INTRODUCTION

Phishing is a form of cybercrime where an attacker posing as a legitimate sender targets an individual or an organization using emails, telephone, or text messages to lure them into providing sensitive, personal, and/or confidential data [1, 4]. This can lead to identity theft, financial and reputational losses [1]. The Anti-Phishing Working Group (APWG) reports that the fourth quarter of 2020 saw an increase in the number of phishing attacks as compared to the first quarter [4], and most of them targeted webmail and performed BECs. The ongoing pandemic-related situation has also given rise to the Covid-19 themed phishing and malware attacks [4]. While phishing detection tools have evolved over the years, so did phishing emails, making it more difficult for tools to detect them [5]. Additionally, attackers often find new methods to circumvent these phishing detection tools [48]. Phishing, a social engineering technique, targets vulnerable individuals to carry out attacks [52]. A previous study [11] shows that individuals are prone to phishing attacks despite their education, gender, hours of computer use, and several other demographic factors.

Proper awareness of phishing attacks and training on how to identify them has been suggested as a method to reduce employees falling for phishing [52, 53]. Training is a proactive method to educate users, thereby enabling them to make decisions on their own [49]. The traditional methods of training using documents, articles, and slideshows are often regarded as non-engaging and monotonous methods [16, 18, 47]. Defined as "games with a purpose", serious games [15] tackle this problem by providing education and engagement. Owing to the positive results shown by serious games [14, 16, 18], the Cybersecurity and Privacy Research Group (CPRG) of our organization took a proactive approach to provide a simulated phishing email awareness training to the employees and study its effects on the participants. The research questions we address are: RQ1) Can a simulated email-based awareness game for enterprise employees show positive effects on their email judgment? RQ2) Can the decision-making steps provide insights on employees' email judgment trends? RQ3) What are the implications that the study can offer in designing email phishing control measures? Thus, we developed *PickMail* to train the employees on how to identify emails with suspicious sender domains, website hyperlinks, fraudulent attachments, and embedded forms. We deployed this game during our organization's annual 'Information Security Awareness Week'. It was played by 478 participants, whose responses were recorded throughout the game. We measured the game's effectiveness by carefully examining the participants' judgments of email legitimacy. The participants' game feedback survey responses suggest that they enjoyed and learned from the game. In this paper, we discuss the *PickMail* game design and mechanics, and the findings from our analysis of the participants' gameplay data.

II. BACKGROUND AND RELATED WORK

The BECs have become one of the most common and expensive cybersecurity threats that target employees of organizations for financial gains. These emails often contain personalized messages that email security systems fail to detect [5]. Spear phishing [6, 7] and Whaling [7] are types of phishing attacks that target

Usable Security and Privacy (USEC) Symposium 2022
28 April 2022, San Diego, CA, USA
ISBN 1-891562-79-7
https://dx.doi.org/10.14722/usec.2022.23059
www.ndss-symposium.org, https://www.usablesecurity.net/USEC/

specific employees within an organization. Here, an attacker poses as a colleague or an HR manager or even the CEO, asking the employee to wire transfer an amount, download malware, or even send personal information [4]. Approaches such as authentication protocols, blacklisting, whitelisting, and prototype email-classification techniques have been suggested to help filter phishing emails [46]. Many of these techniques exhibit limitations. Newer methods include statistical analysis and other content-based phishing filters [46]. However, these methods need to be updated regularly as attackers try to craft new emails to outsmart the existing tools and methods. Several existing tools, which include client-side toolbars, extensions, and plugins that rely on blacklist, heuristics, and other methods to identify phishing content, fail to detect many of the current phishing attacks [9]. While machine learning implementation of phishing tools has shown increased success rates [9, 10], their accuracies vary over various types of phishing tricks. Practical challenges like training these systems are also a limitation [49]. Studies suggest that the advancement of phishing methods over time demands periodic improvements to the phishing-detection tools [10].

The National Cyber Security Centre (NCSC), UK, has put forth guidance on defending organizations from phishing attacks and suggests training as a part of the defense [12]. We believe a combination of awareness training methods and high accuracy phishing detection modules can reduce the users getting phished to a greater extent. Training materials like those from *PhishLine* [19] that focus on domain attacks, Uniform Resource Locator (URL) manipulation, and malicious attachments have been studied. However, a study with a serious game *What.Hack* has been shown to be more effective and engaging than the *PhishLine* materials [18]. Studies have shown that serious games have the potential to be more effective in education than textbooks [20]. Serious games have been used for training in academia [30, 31, 32], teaching cultural heritage [21, 22], training social interactions [23], medical education [24, 25] and healthcare [27], manufacturing education [26], language learning [28], cybersecurity awareness [14, 15, 16, 17, 18], and various other fields. A study on the effectiveness of serious games [29] concluded with a fair amount of evidence to suggest that serious games have positive effects on learning. *Smells Phishy* [36] is an educational board game to raise awareness on online phishing scams, which showed positive results. *Anti-phishing Phil* [14], *Phishy* [16], and *What.Hack* [18] focus on anti-phishing education. *GAP* [17], *Passworld* [43], *PASDJO* [33] train in password security awareness. Other games for cybersecurity education include *Control-Alt-Hack* [34], *CyberCIEGE* [15], *Cyberaware* [35] to name a few. Phishing awareness using serious games has shown that the users have had better performance in identifying phishing URLs after playing the game, compared to other control conditions. The study using *What.Hack* [18] has shown that the game has helped participants learn about phishing emails better than the accompanying control methods. However, the method by which the participants arrived at their final decisions was not dealt with in detail. This motivated us to measure the decision-making steps of the participants for email judgment. Our email simulation game *PickMail* focuses on training the employees on several major aspects of a phishing email, such as the sender's domain (domain of the sender's email address), URLs, attachments, and forms.

III. *PICKMAIL* GAME

We designed *PickMail* to simulate a real-life webmail application used within our organization (cf. Figure 1). The game was preferred over other forms of training because of the increased involvement and engagement levels [16, 18, 47] and positive outcomes serious games have had in the past [14, 16, 18]. Here, we present the design and mechanics of *PickMail.*

A. PickMail: Design and Principles

PickMail incorporates the learning science principle of Reflection [37, 38]. Feedback plays a major role in participants' understanding of the concept. Immediate feedback is provided to the participants' email judgments, which is intended to help them reflect on their learnings and revise misconceptions [38]. Our game design lets the participants follow certain procedural methods to identify the emails displayed on the screen and carefully judge their legitimacy. *PickMail* followed a method of increasing difficulty as the levels progress to promote flow experience [39] and provide gradual learning. We introduced extrinsic motivators such as rewards and recognitions to motivate the participants further, as studies have shown that extrinsic motivators during e-learning make the users feel competent [41]. *PickMail* simulates real-world emails and promotes learning through experience, as described by the experiential learning principle [3].

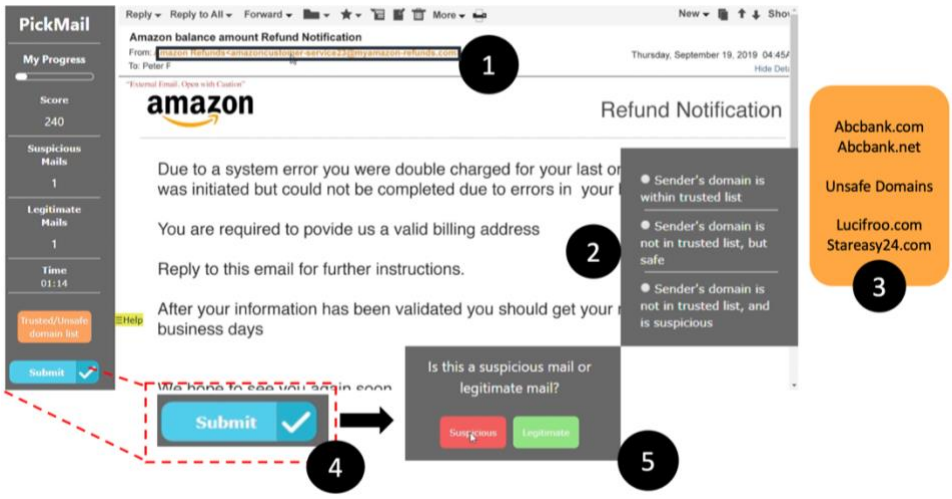

Figure 1: The gameplay of *PickMail*. The player gets to view emails just as they would within the organization's webmail interface. 1) The player can click the parts of email they think are either legitimate or suspicious such as the sender email address domain. 2) A message appears where the player can choose the option that represents the sender domain as "within the trusted list", "not in the trusted list, but safe", and "not in the trusted list, and suspicious". If doubtful, 3) the player can click to view the "trusted/unsafe domain list" to see if the domain is in the list and make a decision appropriately. 4) Once the decision is made, the player can go ahead and click "Submit" button, which then 5) asks to judge if the email is 'Suspicious' or 'Legitimate'. After appropriate judgment, the game will redirect to the page containing the feedback of the email to aid in reflection and better understanding. Note that the features for "Help" and "trusted/unsafe domain list" are not features of the organization's webmail rather they were added to help the participants in the training.

B. PickMail: Game Mechanics

PickMail is a single-player online game developed using HTML and JavaScript. It simulates the process of a virtual email inbox, wherein a player can view the incoming emails and make judgments on their legitimacy by inspecting the various elements within the email. The game takes the role of a story-based agent, which is found to enhance learning experiences [14]. In the game story, the player assumes the role of Peter P, an employee of a fictional organization, who is hoping to get on board to a new project. For this, Peter must pass a test, which involves inspecting and categorizing a set of emails into legitimate and suspicious, that is monitored by the cybersecurity department. Correct responses earn points. Peter must score enough points to clear the threshold (80%) and thus pass the test.

PickMail game starts with a pre-test to check the player's current understanding of phishing emails. Here, a set of four emails are provided for a judgment of legitimacy. This is followed by the gameplay consisting of three levels with an increasing number of elements to identify. Each level has instructions to help the player with the game. Instead of just judging the email, the player must also judge individual elements (at least one) of the email. This is made mandatory throughout the game to let the player understand the importance of careful inspection. The three-level game is followed by a post-test that is used to measure the post-gameplay knowledge of the participants. The post-test is similar to the pre-test but with a different set of emails. The game has 'point-and-click' mechanics where the player can use the mouse to play. Each game level focuses on 1) providing instructions on how to inspect various email elements, 2) measuring participants' decision-making steps, and 3) providing feedback for reflection, thus training participants to judge emails correctly.

1. Providing Instructions on How to Inspect Various Email Elements

The 'Anatomy of a Phishing Email' [44] points to certain methods for identifying phishing emails, such as inspecting the sender's email domain, suspicious website URLs, and attachments. *PickMail* provided instructions on these within each level. The details regarding various phishing emails to be used in the game were provided by our organization's Corporate Security Office (CSO). The in-game emails were modified from several phishing and legitimate emails from our organization's security database. We also referred to certain online instructions from businesses while creating these simulated emails [42, 45, 55]. Level 1 provides instructions on domain-based phishing techniques such as impersonating the sender email from a domain that looks very similar to a real one (for example, modifying ebay.com to eday.com). Level 2 focuses on the use of URLs with word manipulations such as combosquatting [8], wherein the attacker changes the domain names of legitimate URLs by the addition of one or more phrases (for example, twitter.com to twitter-feed.com). The instructions provided to the participants include many types of URLs

that attackers use for luring unsuspecting victims, such as IP address-based URLs (for example, http://173.208.1.73/clickhere.htm), short URLs (for example, bit.ly/s6gTpQ). This level also includes emails that were forwarded multiple times, with the original email containing embedded URLs. Level 3 focuses on malicious attachments (for example, .ZIP and .EXE files) that could install malware when viewed [13, 44]. This level also focused on forms that could collect sensitive user data [44]. The player can view a list of 'Trusted Domains' (cf. Figure 1 (3)) or choose the 'Help' option within a game level for seeking help. The former provides the player with a list of trusted and unsafe domains that help in deciding, and the latter provides tips for the identification of domains, URLs, and attachments based on the level. This method is also employed in the previous study, *What.Hack* [18], which was found to be beneficial to the participants.

2. Measuring Participants' Decision-making Steps

PickMail gauges the participants' decision-making steps needed to inspect and identify various elements of an email before judging its legitimacy. The player can click on any email section that they think are suspicious or legitimate (cf. Figure 1 (1)). This triggers a pop-up (cf. Figure 1 (2)) providing certain choices for the selection. For example, if the selection is for a sender domain (cf. Figure 1 (1)), the options will be to choose if the domain is in the trusted list or not (cf. Figure 1 (2)). Each of the correct decisions gives the player 20 points. While judging all the email elements is not mandatory, the player must inspect and judge at least one of the elements like sender domain, URL, or attachment before judging the email. We believe that such procedural learning will help the participants check every aspect of an email and carefully identify what is legitimate and suspicious before acting on it. After the decisions are made in identifying the elements within the email, the player can go ahead and judge the email legitimacy by clicking the 'Submit' button (cf. Figure 1 (4)). A message is displayed to mark the email as 'legitimate' or 'suspicious' based on the player's judgment.

3. Providing Feedback to Help in Reflection

After the player makes their judgment based on the email presented, immediate feedback is shown to the player regarding the details of the said email. This feedback is displayed in such a way that the player gets an idea of the mistakes made (if any), and it also reinforces the correct judgments, thus helping in reflection. This also educates the player on those elements that were missed. 100 points get added for a correct judgment, whereas 50 points get deducted for a wrong one. We gave this deduction because in real life if a user clicks on a phishing URL and proceeds without caution, there is a chance that it may end up in financial losses. The deduction is a reminder to perform a careful inspection. The player must score at least 80% at the end to pass the game. Only those participants who passed the game were considered for the daily lucky draw for rewards. To support intrinsic experiential learning, we followed the Learning Mechanics – Game Mechanics (LM-GM) Model [50]. The LM-GM model allows the users to relate the learning and game mechanics, and thus reduce the mismatch and suit the game situation to maximize learning. We tried to relate the mechanics of the game story to the learning mechanics of instructional content. Selecting the appropriate email element and providing decisions of whether it is suspicious or legitimate aids in the learning mechanics of task discovery, activity, and problem-solving. The game feedback aids in the reflection and provides motivation. The inclusion of similar elements in multiple emails, such as sender email domains, URLs, and attachments, aids in learning by repetition and reflection.

Table 1: Details of the emails used in Pre-test and Post-test

Pre-test Email Images	Observational points	Post-test Email images	Observational points
Image-1	Sender: 1 Mail body URL:1 Type: Legitimate	Image-1	Sender: 1 Attachment:1 Type: Suspicious
Image-2	Sender: 1 Mail body URL: 0 Type: Suspicious	Image-2	Sender: 2 Mail body URL: 1 Type: Suspicious
Image-3	Sender: 1 Mail body URL: 1 Attachment: 1 Type: Legitimate	Image-3	Sender: 1 Mail body URL: 0 Type: Legitimate
Image-4	Sender: 2 Mail body URL: 1 Type: Legitimate	Image-4	Sender: 2 Mail body URL: 1 Type: Suspicious
Total: 4	9 Observation points	Total: 4	9 Observation points

IV. PICKMAIL: STUDY

A. Study Participants

We conducted an online study and analyzed the results based on participants' responses to in-game emails. The participants of our study were the employees of our organization (IT Sector) who were intimated about *PickMail* through emails. The game was launched as a part of the annual Information Security Awareness Week, and the participants could voluntarily join the game from an internal web portal. We selected lucky winners daily and rewarded them with the company's virtual currency, which translates to actual dollar figure, with which the employees can make purchases. The basic demographic information was collected from the participants using an online survey questionnaire. The details collected include gender, age group, educational and computer science/IT background. The survey was non-mandatory, and the participants were free to skip it as well by responding Not Applicable (NA) to the questions. A total of 478 participants completed *PickMail* game (Gender: [Female: 43.7%, Male: 50.4%, NA: 5.9%]; Age Group: [21-30: 59%, 31-40: 30.7%, 41-50: 4.4%, Over 50: 0.4%, NA: 5.5%]; IT Background: [Yes: 56.5%, No: 38.1%, NA: 5.4%], Educational background: [Bachelor's degree: 66.3%, Master's degree and above: 27%, NA:6.7%]). All the required approvals for the study were obtained from the organization's CSO and the Data Privacy Officer (DPO), who looks after privacy-related matters within the organization. The participants' personal information, such as employee numbers and names, was anonymized so that the game data could not be traced back to any individual. The only information available during analysis was a randomly generated number and the demographic details collected, to ensure that the performance in the game does not negatively affect the participants in any way.

B. Study Method

For our study, we monitored the knowledge levels of our participants using online pre-test and post-test. *PickMail* began with the pre-test (four questions), followed by the game (needs at least 80% to pass), and then the post-test (four questions). The test questions consist of a set of emails that the participants could judge as *legitimate* or *suspicious* based on inspection. Apart from this, the participants could rate their confidence levels on the judgment of each email. This method of identifying games' effectiveness using pre-test and post-test was followed by the previous studies [14, 16, 17, 18]. There were eight emails with an equal number of observational points (Table 1). These emails had sender domains, URLs, attachments, and forward emails to identify.

C. Study Results

1. Analysis of Participants' Judgment Response

We analyzed the effect of *PickMail* by measuring the participants' correctness percentage, false negative and false positive rates, and confidence ratings before and after playing the game. Considering the number of correct answers, we measured a significant improvement in the post-test results ($M = 3.70$, $SD = 0.67$) compared to pre-test results ($M = 3.07$, $SD = 0.86$) ($t(477) = -14.29$, $p<0.0001$, two-tailed paired t-Test).

We measured the False Positive Rate (FPR, when legitimate email is regarded as phishing email) and False Negative Rate (FNR, when phishing email is regarded as legitimate email) using the following formulae:

$$FPR = \text{False Positive (FP)} / (\text{False Positive (FP)} + \text{True Negative (TN)})$$

$$FNR = \text{False Negative (FN)} / (\text{False Negative (FN)} + \text{True Positive (TP)})$$

The latter is riskier as it exposes the user to phishing attacks.

For *PickMail*, the participant data showed a very significant decrease in both FPR and FNR values, with mean FPR decreasing from 0.25 (variance = 0.06) in pre-test to 0.08 (variance = 0.07) in post-test ($t(477) = 10.75$, $p<0.001$, two-tailed paired t-Test) and mean FNR decreasing from 0.17 in pre-test (variance = 0.14)

Table 2: Correctness percentage of the participants based on email type. For email type, "D" denotes the Domain of sender, "U" denotes URLs, and "A" denotes Attachments. The column 'Confidence Ratings' shows the percentage of participants who gave '4' and above as the response to 'Confidence Ratings' in pre-test and post-test (on a scale of 1-5)

Email Type	Correctness Percentage			Confidence Ratings	
Emails having:	Pre-Test (Percentage, variance)	Post-Test (Percentage, variance)	t-Test (t(477))	Pre-Test (Percentage of '4' and above)	Post-Test (Percentage of '4' and above)
Only D	82.42, 0.14	91.84, 0.07	p< 0.01	92.0%	92.5%
D + U	88.70, 0.10	94.97, 0.04	p< 0.01	88.7%	92.3%
D + A	51.67, 0.25	94.97, 0.04	p< 0.01	80.9%	94.6%
Forwarded email	84.51, 0.13	88.70, 0.10	p< 0.1	86.2%	92.7%

to 0.07 (variance = 0.03) in post-test ($t(477) = 6.26$, $p<0.001$, two-tailed paired t-Test). We measured the correctness using the formula:

$$\text{Correctness} = (TP + TN) / (TP + TN + FP + FN)$$

For correctness, there was a significant increase from 76.83% in pre-test (variance = 0.04) to 92.62% (variance = 0.02) in post-test ($t(477) = -14.29$, $p<0.001$, two-tailed paired t-Test). For each question in the pre-test and post-test, we had asked the participants to rate their confidence in answering them as per the 5-point Likert Scale rating [2], from 'least confident' (score = 1) to 'most confident' (score = 5) (with a score of 3 equivalent to 'neutral'). Comparing the pre-test and post-test results, we found an increase in the percentage of participants who gave confidence ratings '4' and '5' in the post-test. The question-wise percentages increased from 88.7% (q1), 92.0% (q2), 80.9% (q3), and 86.2% (q4) in the pre-test to 94.6% (q1), 92.3% (q2), 92.5% (q3), and 92.7% (q4) in the post-test, with average confidence levels being greater in the latter ($z = -10.47$, $p<0.001$, Wilcoxon Signed-Rank Test). Table 2 shows the correctness percentages based on the email types.

Effects on Demographics: We found that males (average correctness increased from 77.1% (pre-test) to 94.5% (post-test), a 22.6% increase, $t(240) = -10.9$, $p<0.001$, two-tailed paired t-Test) showed slightly better correctness increase as compared to females (average correctness increased from 76.7% (pre-test) to 91.7% (post-test), a 19.5% increase, $t(208) = -9.3$, $p<0.001$, two-tailed paired t-Test). The Non-IT background participants showed a significant increase in correctness (from 72.9% in pre-test to 92.5% in post-test, a 26.8% increase, $t(181) = -10.72$, $p<0.001$, two-tailed paired t-Test). This is probably because they had lesser initial knowledge of the topic as compared to the IT-background participants (pre-test: 79.8%, post-test: 93.6%, a 17.3% increase). The age-group of "31-40" showed significant increase in correctness (77.9% to 95.4%, ~22.5%, $t(146) = -8.35$, $p<0.001$, two-tailed paired t-Test), while the 41-50 group showed relatively less percentage increase from pre-test to post-test (83.3% to 95.2%, ~14.3%).

However, their pre-test scores were relatively higher. This might be because the '41-50' age group had more experience in the category of email phishing education as compared to the other groups.

2. Analysis of Participants' Decision-making Steps

We measured the participants' response to the game questions and emails, the number of 'Help' options availed, the timestamp when an email gets displayed (T1), the timestamp when the final decision was made for that email (T2), and the path they followed to reach the conclusions. We also calculated the time spent for each email using (T2 – T1). The order of identification of each element, such as the sender domain, URL, and attachment, was also determined using the timestamps. From Table 3, for emails that had only sender domains to be identified (Email 1 to Email 4), an average of 89.17% of the participants provided correct responses in judging the email. This percentage changed to 83.18% and 88.42%, respectively, for emails where URLs (Email 5 to Email 9) and attachments (Email 11 to Email 13) had to be identified along with sender domains.

Table 3: The emails used within the game and the percentage of participants who gave correct judgments for them, along with their judgments for each email element. The various email elements used to judge legitimacy are sender's domain (the domain of the sender's email address, mentioned as Domain (D) in the Table), URL (U), Form (F), and Attachments (A). "CF" or "Chain forward" specifies the email which has been forwarded multiple times. Forwarded attachment includes those emails which are forwarded, with the original email containing an attachment

Email	Level	Type	Email elements for judgment	Participants with correct email judgments (Out of 478)	Percentage of participants who checked: 'Help' (%)	Domain List (%)	Judged Domain (D) correctly	Judged U/A/F correctly	Judged both D and U/A/F correctly
1		Suspicious	D	438 (91.63%)	25.73	40.38	408	NA	NA
2		Legitimate	D	387 (80.96%)	31.80	56.69	335	NA	NA
3	1	Legitimate	D	463 (96.86%)	28.87	40.38	440	NA	NA
4		Suspicious	D	417 (87.24%)	27.82	39.54	388	NA	NA
5		Suspicious	D+U	408 (85.36%)	20.08	23.64	362	289	257
6		Suspicious	D+U	424 (88.70%)	20.08	16.11	396	260	246
7	2	Suspicious	D+U (CF)	380 (79.50%)	19.87	18.41	252	172	131
8		Legitimate	D+U	367 (76.78%)	20.29	18.83	324	259	234
9		Suspicious	D+U	409 (85.56%)	20.92	25.52	321	314	243
10		Suspicious	D + Form	457 (95.61%)	26.57	6.07	434	299	287
11		Legitimate	D+A	458 (95.82%)	27.41	9.21	426	339	318
12	3	Legitimate	D+A (forward)	401 (83.89%)	22.80	3.77	228	302	211
13		Suspicious	D+A (forward)	409 (85.56%)	25.10	8.16	331	301	271

Email 7 had a chain of forwarded emails, with the original email in the mail trail having a suspicious sender domain and a suspicious URL. 79.5% of participants judged this email to be suspicious. 20.5% failed to identify the initial email in the chain correctly. Considering the average time of judging, the participants took more time to judge forwarded emails, with 50.43 seconds for emails having URLs and 40.00 seconds for emails having attachments. This shows that participants have spent time carefully inspecting these elements rather than quickly coming to conclusions. However, it is also worth noting that the Level 1 emails, focusing on domain names of senders, had a high average time of completion of 47.55 seconds. Certain previous studies have found that an individual spends on average 15-20 seconds per email [40]. The instructions provided upfront to the participants on how to carefully spot various suspicious elements within an email might have triggered them to look further, hence increasing the inspection times. We found that an average of 28% of participants opted 'Help' for emails with just sender domains to inspect, 20.2% of participants asked help for emails with both sender domains and URLs to inspect, and this percentage reached 25.4% for emails having both sender domains and attachments to inspect. This shows that the participants were more careful in identifying emails with masqueraded domain names and unknown domains, supporting our earlier observation of increased time of analysis of these emails. An option to check a list of 'trusted' and 'unsafe' domains was available throughout the game. We found that more participants checked this list for the emails related to domains (44.24%), followed by the emails on URLs (20.50%) and attachments (6.79%). On average, participants took slightly more time to judge legitimate emails (41.6 seconds (SD = 7.2), correctness: 86.8%, average 'help' availed: 26.2%, average 'trusted domains' availed: 25.7%) as compared to suspicious ones (39.3 seconds (SD =6.9), correctness = 87.4%, 'help' =23.2%, 'trusted/untrusted domains' =22.2%).

D. Observations from the Analysis of Participants' Email Judgment

From our analysis, we found that majority of the participants identified multiple email elements correctly before judging the email. The observations are as follows (cf. Figure 2 and Table 3):

1. Most of the participants identified sender domains correctly before making a judgment in level 1.

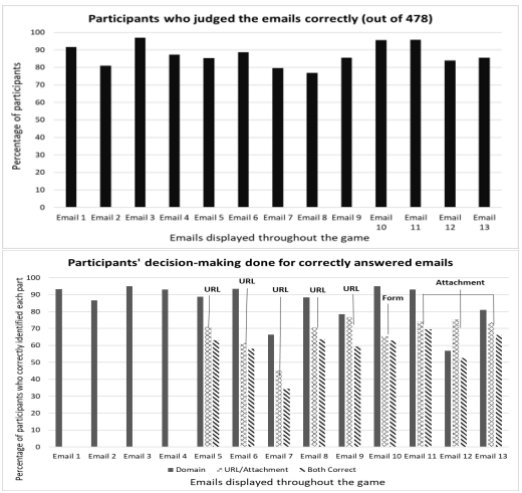

Figure 2 a) Participants' correct decisions throughout the game, b) Participants' judgment based on various components of the email. The Domain, URL/Attachment, and Both Correct are measured as percentages of the correct decisions

2. For forwarded email with attachment (Email 12), out of the 83.9% participants who correctly judged the email, 89% correctly judged the first sender, 56% correctly judged the second (original) sender, 75.3% correctly judged the attachment, with 50.1% participants judging all the elements correctly.
3. The average percentage of participants who checked the 'trusted domain list' came down drastically from Level 1 (44.25%) to Level 3 (6.80%).
4. Irrespective of whether the email is phishing or legitimate, ~61% of participants went on to inspect the second element within an email (URL or attachment).
5. Participants judged multiple elements despite correctly identifying the sender domain as suspicious.
6. An average of 57.82% of participants inspected URLs and/or attachments after inspecting the sender's email address domain.
7. Some participants made false positive judgments (23.2% for Email 8). This could also mean that the participants were more careful with the said email and decided to mark it as suspicious. In real life, this is indeed less risky than incorrectly judging a phishing email as legitimate.
8. The 'help' options availed by the participants were evenly distributed among the levels, with 95.2% of participants asking 'help' once per level. This shows that once participants learned the judgment methods, they made their decisions by themselves.
9. With 'help' opted, the average time per email was 81.98 seconds (correctness = 88.0%). For emails where no 'help' opted, the average time per email was found to be 27.4 seconds (correctness = 86.9%).

E. Comparison with the Existing Game-based Study on Phishing Education Based on Reported Values

Comparing *PickMail's* correctness percentages to those of the previous game *What.Hack* [18], *PickMail* resulted in a correctness increase from 76.8% in the pre-test to 92.6% in the post-test. The correctness percentage of *What.Hack* reached 89% (from 65%). The correctness values in the pre-test are relatively higher than the *What.Hack*. We believe that this is a result of the demographic (IT crowd) being mostly familiar with various cybersecurity-related topics and almost everyone has had some sort of training in the past when it comes to cybersecurity awareness.

Unlike in *What.Hack*, *PickMail* focuses on participants' decision-making steps and not only on their final email judgments. The game is influenced by the organization's webmail interface, and it includes the organization's emails (duly modified for the game) to make the participants experience the game as close to real-life as possible. *What.Hack* focuses on a small crowd; however, *PickMail* scaled up the study to a large enterprise audience. The overall count of game questions in *PickMail* is lesser as compared to *What.Hack*. *PickMail* was initially designed to have 20 questions (emails), divided into five levels. However, unlike in *What.Hack*, where 20 emails are subjected to binary decision-making, in *PickMail*, the player has to make two to four (average greater than 3) decisions per email. This meant that the overall playtime for the 5-level game was coming out to be more than 30 minutes, which was not desirable in the enterprise setting. Therefore, we modified *PickMail* to have three levels and 13 emails. While *PickMail* showed higher correctness percentages, our demographics are also different as compared to the other games, which could have helped in getting better results.

F. Participants' Feedback about PickMail

We asked the participants about their engagement in the game and their understanding of the concepts. We measured their feedback based on three questions on fun, educational ability, and learnability. A total of 444 participants provided the game feedback survey (which was voluntary). The Likert Scale [2] of responses for the question vary from 'strongly disagree' (a value of 1) to 'strongly agree' (a value of 5), with the central value being 'neutral' (a value of 3). We found that 83.10% participants found the game to be fun (M = 4.11, variance = 1.20), 84.68% found the game to be educational (M = 4.22, variance = 1.25), and 83.55% participants responded that they learned about phishing email identification through the game (M = 4.13, variance = 1.28). Apart from these, we also received various feedback comments from the participants such as "Good Quiz and very informative. The best part is that it provided the answers then and there to understand what we have missed", "Really good Gamification of the topic. Thoroughly enjoyed it.", "Great experience. If same is included in security learnings, it will be very helpful for the associates".

V. DISCUSSIONS

To answer our RQ1, the participants showed improvements in correctness percentages after playing the game, as shown by the post-test results. The correctness in the post-test increased to 92.6%, from 76.8% in the pre-test. We believe this increase is because of how they treated the emails based on their learnings and feedback from the game. The Non-IT crowd has shown relatively better responses, which could suggest that future training should focus more on users with lesser exposure to computer-IT knowledge. In the current scenario where internet connectivity is increasing daily, a similar kind of training could benefit the general populace. From the game feedback survey, we also found that the participants enjoyed the game.

Considering the decision-making steps taken, we found that most of the participants correctly judged email's sender domains irrespective of whether the emails had URLs or attachments. This shows that the participants learned to check for the source of the email. To answer RQ2, we found that the enterprise employees judged the emails based on individual inspection of email elements. This is a positive finding, suggesting that the employees follow a practice that helps them break down an incoming email to separate the legitimate from the phishing. This is also in line with the findings of Wash [51] on the cognitive processes by which IT experts identify phishing emails.

Participants spent slightly more time judging legitimate emails, for which they also opted for more 'Help' options. This could be likely because once they judged the sender domain as legitimate, they searched the email further to find at least one suspicious element. Some participants judged legitimate emails as 'suspicious', thus making false-positive judgments. While this leads to incorrect answers within the game, it can be considered as a precaution. For legitimate emails, when the player has identified the sender domain to be legitimate, they would have checked for any suspicious elements within the email. A small number of participants judged the legitimate emails to be 'suspicious'. Unless the emails are confirmed to be from known senders, taking precautions and ignoring the email might protect the user from an unknown threat. However, such email domains should also be subjected to proper evaluation and scrutiny so that the false positives get minimized over time. Facilities within the email clients for whitelisting legitimate URLs could help the users. There is a need for a balance between

caution and action while dealing with emails, which future training methods could focus on.

A. Implications to Reduce Users Falling for Phishing Attacks

1. More focus on Experiential Learning: The participants' high correctness percentages, decision-making analysis, and feedback responses emphasize the importance of providing experiential training [3]. We believe that this training methodology could be used in multiple contexts such as password creation, classification of documents, user consent, and privacy-related topics and could be a focus of future research.

2. Training Methods should focus on using user-specific content for awareness: We believe that the game provided a familiar experience to the participants in viewing and judging emails. This could also have helped them identify the email elements quicker within the game. The newer games and training methods in this domain should have learning content that could be populated with contextual data (for example, enterprise emails) for providing a realistic experience to the participants, and thus have a more impactful training.

3. Feedback plays an important role: The game provides both instructions and feedback for communicating the learning content. The instructions are also available as 'Help' option, and we found that participants visited 'Help' evenly throughout the game. The inspection of email elements within levels also suggests that the participants have read the feedback that suggests them to do so. Studies show feedback helps in reflection [37, 38]. Training methods on phishing awareness should focus on incorporating sufficient and clear information to ensure that the participants receive appropriate feedback for their responses.

4. Smart email client user interface: Participants spent more time judging emails within the game as compared to the time reported by previous studies, more likely because they tried to identify various suspicious elements within the emails. This could also be motivated by extrinsic motivators such as points and rewards. However, this is less likely, as seen from the game feedback survey and the fact that the reward was lottery-based. We believe the time required for email inspection could be greatly reduced if the email client interfaces provide support to the users in email judgments. Providing emphasis to embedded URLs, attachments, and suspicious sender domains could be an initial step. Facilities to whitelist known domains, expand short URLs within emails, and provide actual redirection addresses of URLs could be the next step(s) that email clients could take to help the users in proper email judgment, thus reducing the likelihood of phishing.

B. Limitations and Future Scope

For *PickMail*, we did not exclusively carry out a control condition. We intended to scale up the previous study [18] for orders of magnitude larger audiences in an enterprise setting and analyze the participants' decision-making steps. We also plan to launch the game to a wider audience, especially to the non-IT crowd, in the future. The long-term effects of our training are yet to be studied. Since newer phishing attacks are appearing daily, training on these updated practices will be necessary. To ensure that the training has equipped the participants for better decision-making, we are also planning a new study on how partial training affects users. For example, if we were to train users on spotting various ambiguities in sender domains in emails, and then provide them with emails having URLs (legitimate and suspicious), how would the users react? Will they just judge the email based on the learning they had about the sender domains, or will they try to look further? We believe that an extensive game on training users on additional types of attachments and URLs within emails could be more beneficial in the future.

VI. CONCLUSIONS

We developed an email phishing awareness training game *PickMail* to train our organization's employees on how to judge a given email to be a phish or not. We trained the participants on judging email legitimacy by inspecting various parts of the email such as sender domain, URLs, attachments, and forms. The answers to our research questions were found from the analysis of the participants' responses. From the post-game test data from 478 participants, we observed 92.62% correctness in judging emails, an increase from 76.8% in the pre-game test, suggesting positive results on the email judgment. The FNR (0.07) and FPR (0.08) values were lower compared to similar games for training phishing awareness. From the data on participants' steps taken for email judgment, we found that most of them correctly judged the emails based on their assessments of individual elements of the email. While the participants spent more time judging emails, they also carefully inspected the individual elements. The implications of our study could help in designing better learning content for email-phishing awareness and designing better user interfaces. While our study showed a glimpse of the effectiveness of a serious game-based phishing awareness, widening the scope of the learning content within user-specific contexts could help users experience and understand the necessary steps required to combat phishing.

REFERENCES

[1] Phishing.org. “What is Phishing?” Retrieved June 07, 2021 from https://www.phishing.org/what-is-phishing

[2] I. Elaine Allen and Christopher A. Seaman. 2007. Likert Scales and Data Analyses. July 2007. Retrieved February 19, 2021 from http://rube.asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html

[3] Kolb, D. A. (2014). Experiential learning: Experience as the source of learning and development. FT press.

[4] Anti Phishing Working Group. 2021. Phishing Activity Trends Reports. Retrieved June 07, 2021 from https://apwg.org/trendsreports/

[5] Cidon, Asaf, et al. "High precision detection of business email compromise." 28th {USENIX} Security Symposium ({USENIX} Security 19). 2019.

[6] Parmar, Bimal. "Protecting against spear-phishing." Computer Fraud & Security 2012.1 (2012): 8-11.

[7] Pienta, Daniel, Jason Bennett Thatcher, and Allen C. Johnston. "A taxonomy of phishing: Attack types spanning economic, temporal, breadth, and target boundaries." Proceedings of the 13th Pre-ICIS Workshop on In-formation Security and Privacy, San Francisco, CA, USA. Vol. 1. 2018.

[8] Kintis, P., Miramirkhani, N., Lever, C., Chen, Y., Romero-Gómez, R., Pitropakis, N., ... & Antonakakis, M. (2017, October). Hiding in plain sight: A longitudinal study of combosquatting abuse. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 569-586).

[9] Coffie, I., Njogu, G. N., Mabarani, S., Rubaiza, J. B., & Svotwa, L. Enhancing Phishing Detection Tools: A Machine Learning Approach.

[10] Viktorov, O. (2017). Detecting phishing emails using machine learning techniques (Doctoral dissertation, Middle East University).

[11] Dhamija, R., Tygar, J. D., & Hearst, M. (2006, April). Why phishing works. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 581-590).

[12] National Cyber Security Centre. 2018. Phishing attacks: defending your organization. Retrieved June 10, 2021 from https://www.ncsc.gov.uk/guidance/phishing

[13] Hong, J. (2012). The state of phishing attacks. Communications of the ACM, 55(1), 74-81.

[14] Sheng, Steve, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. "Anti-phishing Phil: the design and evaluation of a game that teaches people not to fall for phish." In Proceedings of the 3rd symposium on Usable privacy and security, pp. 88-99. ACM, 2007.

[15] Thompson, Michael, and Cynthia Irvine. "Active learning with the CyberCIEGE video game." (2011)

[16] CJ, Gokul., Pandit, S., Vaddepalli, S., Tupsamudre, H., Banahatti, V., & Lodha, S. (2018, October). Phishy-a serious game to train enterprise users on phishing awareness. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts (pp. 169-181).

[17] Tupsamudre, Harshal, Rahul Wasnik, Shubhankar Biswas, Sankalp Pandit, Sukanya Vaddepalli, Aishwarya Shinde, C. J. Gokul, Vijayanand Banahatti, and Sachin Lodha. "GAP: A Game for Improving Awareness About Passwords." In Joint International Conference on Serious Games, pp. 66-78. Springer, Cham, 2018.

[18] Wen, Zikai Alex, et al. "What. hack: engaging anti-phishing training through a role-playing phishing simulation game." Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019.

[19] 2018. PhishLine Training. https://www.phishline.com/complimentary-content/

[20] Stege, L., Van Lankveld, G., & Spronck, P. (2011). Serious games in education. International Journal of Computer Science in Sport, 10(1), 1-9.

[21] Mortara, M., Catalano, C. E., Bellotti, F., Fiucci, G., Houry-Panchetti, M., & Petridis, P. (2014). Learning cultural heritage by serious games. Journal of Cultural Herit-age, 15(3), 318-325.

[22] Bampatzia, S., Bourlakos, I., Antoniou, A., Vassilakis, C., Lepouras, G., & Wallace, M. (2016, December). Serious games: valuable tools for cultural heritage. In International Conference on Games and Learning Alliance (pp. 331-341). Springer, Cham.

[23] Grossard, C., Grynspan, O., Serret, S., Jouen, A. L., Bailly, K., & Cohen, D. (2017). Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (ASD). Computers & Education, 113, 195-211.

[24] Graafland, M., Schraagen, J. M. C., & Schijven, M. P. (2012). Systematic review of validity of serious games for medical education and surgical skills training. The British journal of surgery, 99(10), 1322-30.

[25] Gorbanev, I., Agudelo-Londoño, S., González, R. A., Cortes, A., Pomares, A., Delgadillo, V., ... & Muñoz, Ó. (2018). A systematic review of serious games in medical education: quality of evidence and pedagogical strategy. Medical education online, 23(1), 1438718.

[26] Pourabdollahian, B., Taisch, M., & Kerga, E. (2012). Serious games in manufacturing education: Evaluation of learners’ engagement. Procedia Computer Science, 15, 256-265.

[27] Arnab, S., Dunwell, I., & Debattista, K. (2013). Serious games for healthcare: Applications and implications. Medical Information Science Reference.

[28] Sørensen, B. H., & Meyer, B. (2007). Serious Games in language learning and teaching-a theoretical perspective. In DiGRA Conference (pp. 559-566).

[29] Backlund, P., & Hendrix, M. (2013, September). Educational games-are they worth the effort? A literature survey of the effectiveness of serious games. In 2013 5th international conference on games and virtual worlds for serious applications (VS-GAMES) (pp. 1-8). IEEE.

[30] Prensky, M. (2003). Digital game-based learning. Computers in Entertainment (CIE), 1(1), 21-21.

[31] Huizenga, J., Admiraal, W., Akkerman, S., & Dam, G. T. (2009). Mobile game-based learning in secondary education: engagement, motivation and learning in a mobile city game. Journal of Computer Assisted Learning, 25(4), 332-344.

[32] Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. Computers in human behavior, 54, 170-179.

[33] Seitz, Tobias, and Heinrich Hussmann. "PASDJO: quantifying password strength perceptions with an online game." Proceedings of the 29th Australian Conference on Computer-Human Interaction. ACM, 2017.

[34] Denning, Tamara, Adam Lerner, Adam Shostack, and Tadayoshi Kohno. "Control-Alt-Hack: the design and evaluation of a card game for computer security awareness and education." In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, pp. 915-928. ACM, 2013.

[35] Giannakas, Filippos, Georgios Kambourakis, and Stefanos Gritzalis. "Cyberaware: A mobile game-based app for cybersecurity education and awareness." Interactive Mobile Communication Technologies and Learning (IMCL), 2015 International Conference on. IEEE. 2015.

[36] Baslyman, M., & Chiasson, S. (2016, June). " Smells phishy?": An educational game about online phishing scams. In 2016 APWG Symposium on Electronic Crime Research (eCrime) (pp. 1-11). IEEE.

[37] Donovan, M. Suzanne, John D. Bransford, and James W. Pellegrino. How people learn: Bridging research and practice. National Academy Press, 2101 Constitution Avenue NW, Lockbox 285, Washington, DC 20055, 1999.

[38] Penuel, B., Roschelle, J., & Cohen, A. L. (1999). Designing learning: Cognitive science principles for the innovative organization. Designing learning: Principles and technologies (SRI paper series). SRI Project, 10099.

[39] Csikszentmihalyi, Mihaly, Sami Abuhamdeh, and Jeanne Nakamura. "Flow." Flow and the foundations of positive psychology. Springer, Dordrecht, 2014. 227-238.

[40] Marketing Sherpa. Alarming Research Results. January 13, 2005. Retrieved June 10, 2021 from https://www.marketingsherpa.com/article/average-email-open-time-is#

[41] Yoo, S. J., Han, S. H., & Huang, W. (2012). The roles of intrinsic motivators and extrinsic motivators in promoting e-learning in the workplace: A case from South Korea. Computers in Human Behavior, 28(3), 942-950.

[42] Amazon.com. Internet scams and phishing. Retrieved December 18, 2020 from https://pay.amazon.com/help/201754760

[43] Jayakrishnan, G. C., Sirigireddy, G. R., Vaddepalli, S., Banahatti, V., Lodha, S. P., & Pandit, S. S. (2020). Password: A Serious Game to Promote Password Awareness and Diversity in an Enterprise. In Sixteenth Symposium on Usable Privacy and Security ({SOUPS} 2020) (pp. 1-18).

[44] Drake, C. E., Oliver, J. J., & Koontz, E. J. (2004, August). Anatomy of a Phishing Email. In CEAS.

[45] Amazon. Seller Forums. Retrieved December 10, 2020 from https://sellercentral.amazon.com/forums/

[46] Bergholz, A., De Beer, J., Glahn, S., Moens, M. F., Paaß, G., & Strobel, S. (2010). New filtering approaches for phishing email. Journal of computer security, 18(1), 7-35.

[47] Cone, Benjamin D., Cynthia E. Irvine, Michael F. Thompson, and Thuy D. Nguyen. "A video game for cyber security training and awareness." computers & security 26, no. 1 (2007): 63-72.

[48] Gupta BB, Tewari A, Jain AK, Agrawal DP (2017) Fighting against phishing attacks: state of the art and future challenges. Neural Comput Appl 28(12):3629–3654

[49] Jampen, D., Gür, G., Sutter, T., & Tellenbach, B. (2020). Don't click: towards an effective anti-phishing training. A comparative literature review. Human-centric Computing and Information Sciences, 10(1), 1-41.

[50] Lim, T., Carvalho, M. B., Bellotti, F., Arnab, S., De Freitas, S., Louchart, S., ... & De Gloria, A. (2015). The LM-GM framework for serious games analysis. Pittsburgh: University of Pittsburgh.

[51] Rick Wash. 2020. How Experts Detect Phishing Scam Emails. Proc. ACM Hum.-Comput. Interact. 4, CSCW2, Article 160 (October 2020), 28 pages. DOI:https://doi.org/10.1145/3415231

[52] Monk, Thomas, Johan Van Niekerk, and Rossouw Von Solms. "Concealing the Medicine: Information Security Education through Game Play." ISSA.2009.

[53] Mayhorn, Christopher B., and Patrick G. Nyeste. "Training users to counteract phishing." Work 41.Supplement 1 (2012): 3549-3552.

[54] Jampen, Daniel, et al. "Don't click: towards an effective anti-phishing training. A comparative literature review." Human-centric Computing and Information Sciences 10.1 (2020): 1-41.

[55] Stay Safe From Scammers. eBay Security Center. Retrieved from https://pages.ebay.com/security center/stay_safe.html

APPENDIX

A SUPPLEMENTARY MATERIAL: QUESTIONS

The following sections show the emails used in the pre-test, post-test (in redacted form), and some screenshots from the game.

A.1 Pre-test Questions

The pre-test questions had options to choose whether the displayed emails were legitimate or suspicious as per the participants' understanding. The questions also provided the option to choose their level of confidence in their responses. This confidence was captured using Likert-scale ratings from one to five, where 'one' refers to 'least confident' and 'five' refers to 'very confident'. The same methodology was used for post-test questions

as well. The pop-up question that is asked along with a pre-test/post-test email is given below, followed by the pre-test questions.

Figure A1. Pop-up asked along with a pre-test/post-test email

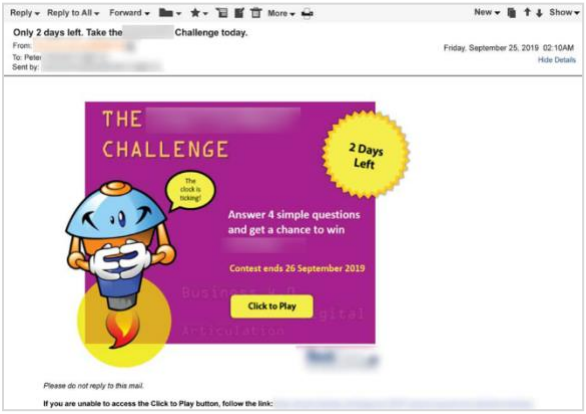

Figure A2. Pre-test Question 1

.

Figure A3. Pre-test Question 2

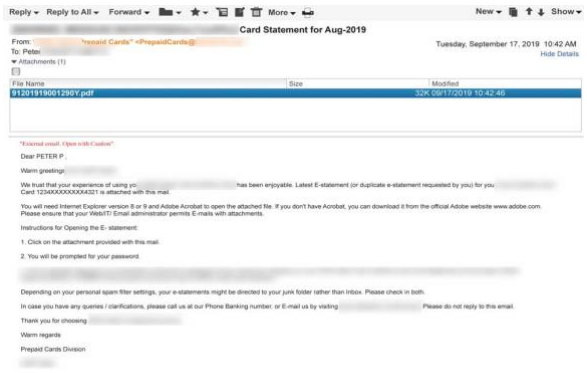

Figure A4. Pre-test Question 3

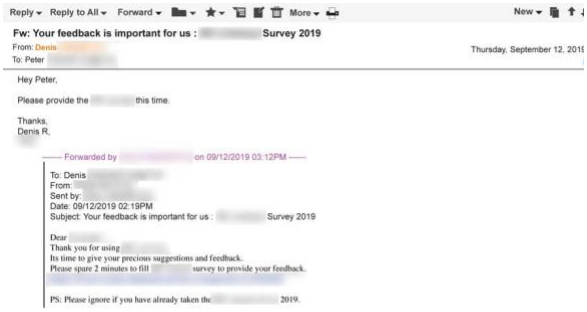

Figure A5. Pre-test Question 4

A.2 Post-test Questions

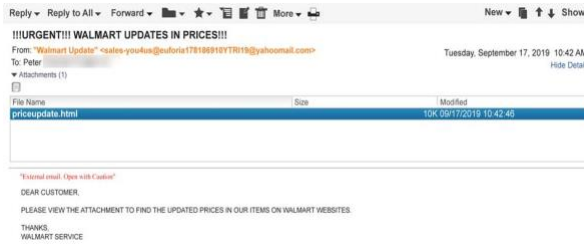

Figure A6. Post-test Question 1

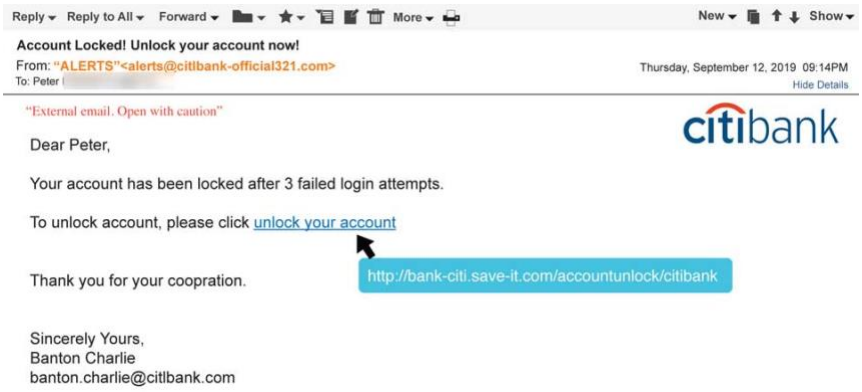

Figure A7. Post-test Question 2

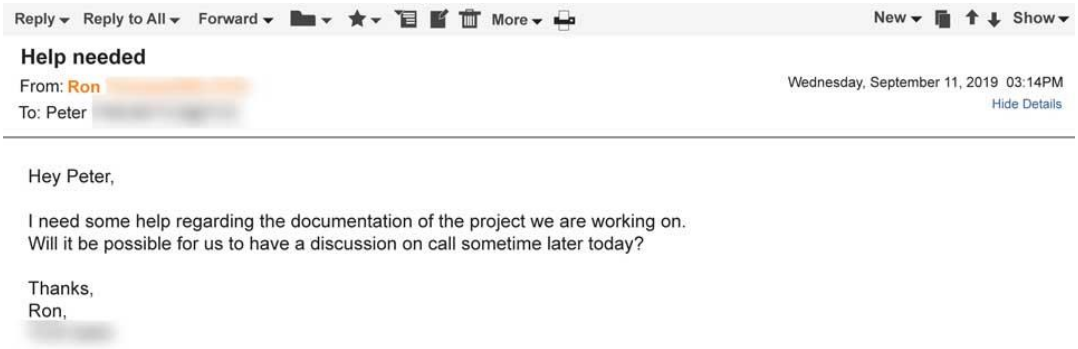

Figure A8. Post-test Question 3

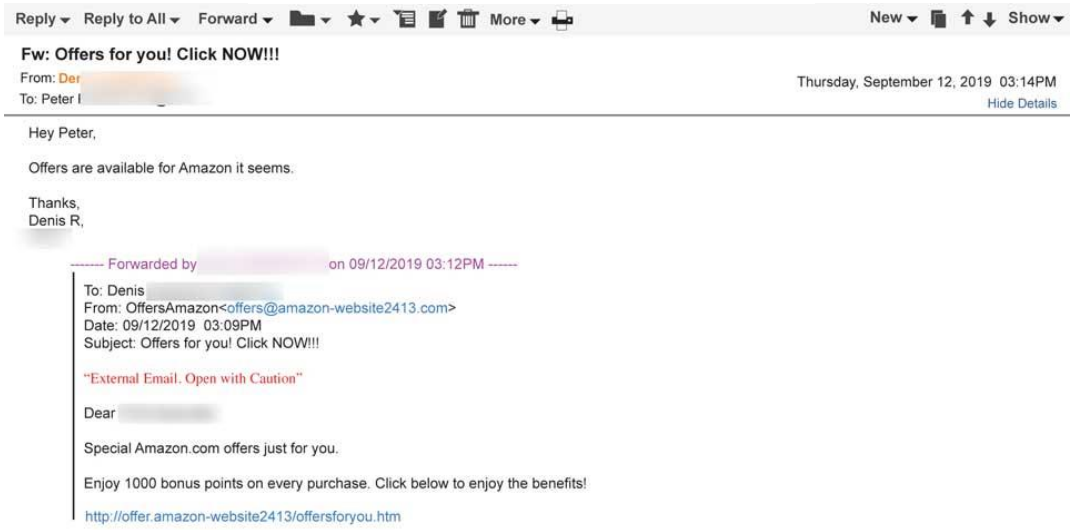

Figure A9. Post-test Question 4