Getting Users to Pay Attention to Anti-Phishing Education: Evaluation of Retention and Transfer

Ponnurangam Kumaraguru, Yong Rhee, Steve Sheng, Sharique Hasan,
Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong

Carnegie Mellon University

ponguru@cs.cmu.edu, ywrhee@cmu.edu, shengx@cmu.edu, shasan@andrew.cmu.edu,
acquisti@andrew.cmu.edu, lorrie@cs.cmu.edu, jasonh@cs.cmu.edu

ABSTRACT

Educational materials designed to teach users not to fall for phishing attacks are widely available but are often ignored by users. In this paper, we extend an embedded training methodology using learning science principles in which phishing education is made part of a primary task for users. The goal is to motivate users to pay attention to the training materials. In embedded training, users are sent simulated phishing attacks and trained after they fall for the attacks. Prior studies tested users immediately after training and demonstrated that embedded training improved users' ability to identify phishing emails and websites. In the present study, we tested users to determine how well they retained knowledge gained through embedded training and how well they transferred this knowledge to identify other types of phishing emails. We also compared the effectiveness of the same training materials delivered via embedded training and delivered as regular email messages. In our experiments, we found that: (a) users learn more effectively when the training materials are presented after users fall for the attack (embedded) than when the same training materials are sent by email (non-embedded); (b) users retain and transfer more knowledge after embedded training than after non-embedded training; and (c) users with higher Cognitive Reflection Test (CRT) scores are more likely than users with lower CRT scores to click on the links in the phishing emails from companies with which they have no account.

Categories and Subject Descriptors

D.4.6 Security and protection, H.1.2 User / Machine systems, H.5.2 User interfaces, K.6.5 Security and protection education.

General Terms

Design, Experimentation, Security, Human factors.

Keywords

Embedded training, learning science, instructional principles, phishing, email, usable privacy and security, situated learning.

1. INTRODUCTION

Users are susceptible to phishing attacks because of the sensitive trust decisions that they make when they conduct activities online.

Psychologists have shown that people do not reflect on their options when making decisions under stress (e.g. accessing email while busy at work). Studies have shown that people under stress fail to consider all possible solutions and may end up making decisions that are irrational [11]. Psychologists call this the *singular evaluation approach* to decision making. In this approach, people evaluate solution options individually rather than comparing them with others, taking the first solution that works [13, pp. 20]. Psychologists have also shown that people do not ask the right questions when making decisions under stress and also rely on familiar patterns instead of considering all relevant details [28, 29].

Anti-phishing researchers have developed several approaches to preventing and detecting phishing attacks [9, 24], and to supporting Internet users in making better trust decisions that will help them avoid falling for phishing attacks. Much work has focused on helping users identify phishing web sites [8, 25, 26]. Less effort has been devoted to developing methods to train users to be less susceptible to phishing attacks [15, 20, 23].

Researchers argue that user education in the context of security is difficult because (1) security is always a secondary task for the end-users [30], (2) users are not motivated to read about privacy and security [4], and (3) users who do read about privacy and security develop a fear of online transactions, but do not necessarily learn how to protect themselves [1]. However, our hypothesis - which was validated through experiments - is that people can be taught to identify phishing scams without necessarily understanding complicated computer security concepts [15].

In this paper, we extend an embedded training methodology using learning science principles in which phishing education is made part of a primary task for users. The goal is to motivate users to pay attention to the training materials. In embedded training, users are sent simulated phishing attacks and are presented training interventions if they fall for the attacks. Prior studies tested users immediately after training and demonstrated that embedded training improved users' ability to identify phishing emails and websites. They also compared embedded training to security notices delivered via email. However, the security notices did not include the same content as the embedded training materials [15]. In the present study, we tested users to determine how well they retained knowledge gained through embedded training over a period of about one week, and how well they transferred this knowledge to identify other types of phishing emails. We also compared the effectiveness of the same training materials

delivered via embedded training and delivered as a regular email message (non-embedded). In our experiments, we found that: (a) users learn more effectively when the training materials are presented after users fall for the attack (embedded) than when the same training materials are sent by email (non-embedded); (b) users retain and transfer more knowledge after embedded training than after non-embedded training; and (c) users with higher Cognitive Reflection Test (CRT) scores are more likely than users with lower CRT scores to click on the links in the phishing emails from companies with which they have no account.

The remainder of the paper is organized as follows: In the next section we describe the embedded training methodology, and learning science principles that we used in developing the training methodology. In Section 3, we present the theory and the hypotheses that guided our study. In Section 4, we present the study methodology used to test our hypotheses. In Section 5, we present the results of our evaluation, demonstrating that embedded training is more effective than non-embedded training, and that users can retain over time and transfer the knowledge gained. We discuss the effect of training users in Section 6. Finally, we present our conclusions and future work in Section 7.

2. TRAINING

In this section we describe the background of the embedded training concept and how the methodology works. We also describe the learning science principles that we applied while designing the embedded methodology and the training materials.

2.1 Embedded training

Education researchers argue that training is most effective if the training materials incorporate the context of the real world, work, or testing situation [3]. Embedded training is a methodology in which training materials are integrated into the primary tasks that users perform in their day-to-day lives. Researchers define embedded training as the ability to train a task using the associated operational system including software and machines that people normally use. This training methodology has been widely applied to the training of military personnel on new Future Combating Systems (FCS) [12].

In our application of embedded training for protecting people from phishing, we send users simulated phishing emails that urge them to click on a link to visit a web site, login, and provide personal information. In a deployed embedded training system, these emails might be sent by a corporate system administrator, ISP, or training company. We intervene and show training materials to users when they click on links in the email. We consider this to be the point where most people fall for the phish because evidence from laboratory studies suggests that most users who click on links in phishing emails go on to provide their personal information to phishing web sites. For example, in one lab study 93% of the participants who clicked on links provided their personal information [15]. Our training intervention messages explain to users that they are at risk for phishing attacks and give them tips for protecting themselves (as shown in Figure 1). This approach has the following advantages: (1) it enables a system administrator or training company to continuously train
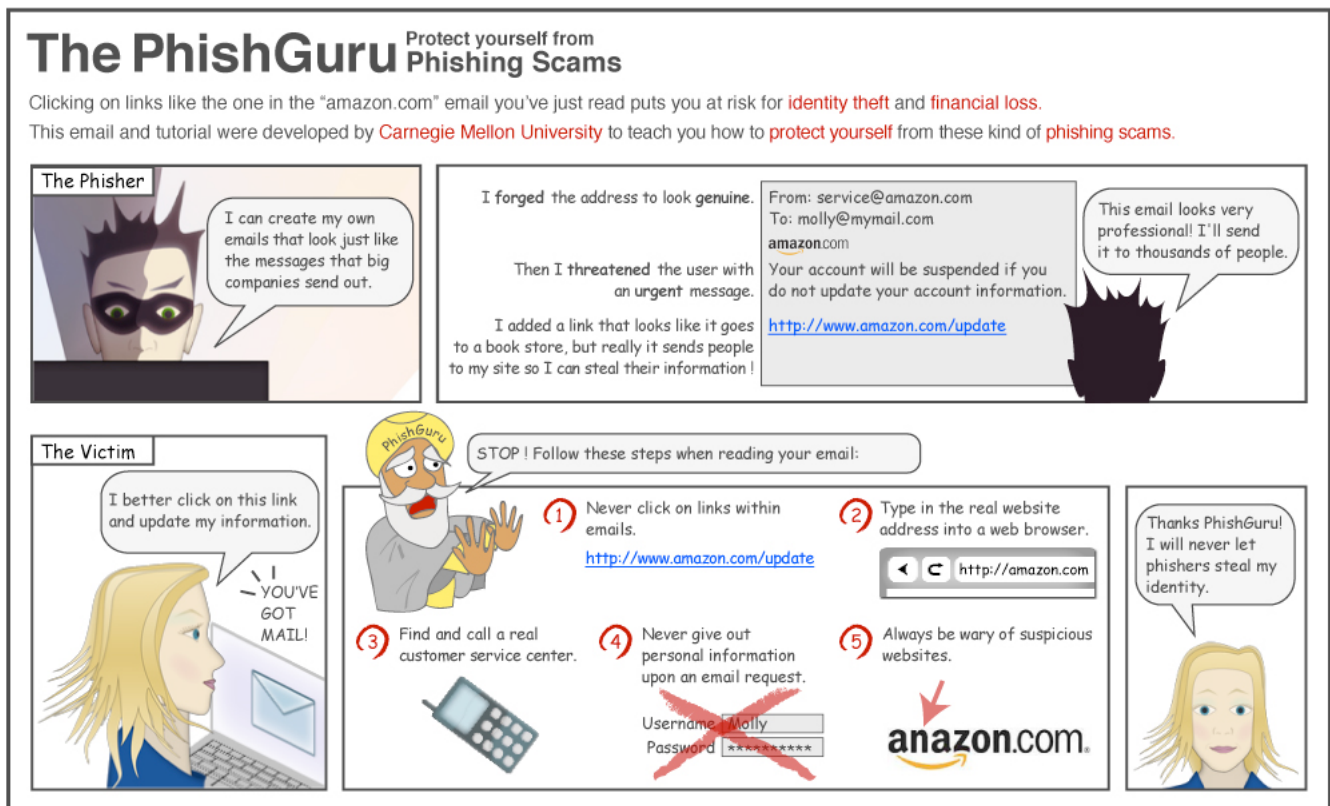

Figure 1. The comic strip intervention uses a comic strip to tell a story about how phishing works and how people can protect themselves.

people as new phishing methods arise; (2) it enables users to get trained without taking time out of their busy schedules (making the training part of primary task); and (3) it creates a stronger motivation for users because training materials are presented only after they actually fall for a phishing email.

2.2 Learning science principles

Anandpara et al. have shown that users who read existing online training materials become concerned about phishing and tend to become overly cautious, identifying many legitimate sites as fraudulent [1]. However, they do not learn useful techniques for identifying phishing emails and web sites because much of the existing online training does not teach specific cues and strategies [14]. In addition, it appears that most of existing online training materials were not designed using instructional design principles from learning science. To maximize the effectiveness of our training materials, we applied the following learning science principles:

- *Learning-by-doing*: One of the fundamental hypotheses of Adaptive Control of Thought–Rational (ACT-R) theory of cognition and learning is that knowledge and skills are acquired and strengthened through practice (by doing) [2]. In our approach, users learn by actually clicking on phishing emails (doing). The training materials are presented when users fall for phishing emails.

- *Immediate feedback*: Researchers have shown that providing immediate feedback during the learning phase results in more efficient learning and faster learning, provides guidance towards correct behavior, and reduces unproductive floundering [22]. In our approach, we provide feedback through interventions immediately after the user clicks on a link in a fake phishing email sent by us.

- *Contiguity principle*: Mayer et al. developed the contiguity principle, which states that "the effectiveness of the computer aided instruction increases when words and pictures are presented contiguously (rather than isolated from one another) in time and space" [17]. In our design, we have placed pictures and relevant text contiguously in the instructions (numbered 1 through 5 in Figure 1). For example, in Figure 1, instruction 2, we present the image of a browser in order to convey the instruction "Type in the real website address into a web browser."

- *Personalization principle*: This principle suggests that "using conversational style rather than formal style enhances learning" [5, Chapter 8]. People make efforts to understand the instructional materials if it is presented in a way that makes them feel that they are in a conversation. The principle recommends using "I," "we," "me," "my," "you," and "your" in the instructional materials to enhance learning [5, 16]. We apply this principle in the design in many ways, for example, "STOP! Follow these steps when reading your email" (Figure 1).

- *Story-based agent environment principle*: Agents are characters who help in guiding the users through the learning process. These characters can be represented visually or verbally and can be cartoon-like or real life characters. The story-based agent environment principle states that "using agents in a story-based content enhances user learning" [19]. People tend to put in efforts to understand the materials if there is an agent who is guiding them in the learning process.

Learning is further enhanced if the materials are presented within the context of a story [16]. People learn from stories because stories organize events in a meaningful framework and tend to stimulate the cognitive process of the reader [13, Chapter 11]. We have created three characters: "The phisher," "The victim," and "The PhishGuru." We applied the story-based agent environment principle by having the top layer (in Figure 1) of the comic script showing what the bad guy (phisher) can do and the bottom layer showing instructions that potential victims can learn to protect themselves from falling for phishing emails.

3. THEORY AND HYPOTHESES

In this section we introduce five hypotheses for the study described in this paper. Three hypotheses relate to user learning and two hypotheses relate to users' susceptibility to phishing emails.

3.1 Learning

Motivation is one of the most important aspects of the learning process. Researchers have shown that users can be trained through an embedded training methodology, where the training is made part of the primary task and users are motivated to learn because they are presented with training materials after falling for phishing emails [15]. However, while that study suggested that embedded training increased the motivation to learn, it did not evaluate if the embedded training approach was necessarily better than sending the same training materials directly via email.

To test the value of embedded training, the present study included three conditions: "embedded," "non-embedded," and "control." Participants in the embedded condition received a simulated phishing email and saw the training materials when they clicked on a link in that email. Participants in the non-embedded condition received the same training materials directly as part of an email message. Participants in the control condition received an additional email from a friend, but received no training.

Hypothesis 1: Users learn more effectively when training materials are presented after they fall for a phishing attack (embedded) than when the training materials are sent by email (non-embedded).

3.2 Retention

A large body of literature focuses on quantifying knowledge retention [21]. Learning science literature defines retention as the ability of learners to retain or recall the concepts and procedures taught when tested under the same or similar situations after a time period δ from the time of knowledge acquisition. Researchers have frequently debated the optimum δ to measure retention [18]. Prior studies that demonstrated that users can be taught to avoid phishing attacks tested users immediately after they were trained and thus did not explore users' ability to retain this knowledge [15, 23]. Thus, the question remains as to whether users retain the knowledge that they have gained during training.

Hypothesis 2: Users retain more knowledge about how to avoid phishing attacks when trained with embedded training than with non-embedded training.

3.3 Transfer

Transfer is the ability to transfer the knowledge gained from one situation to another situation after a time period δ from the time of knowledge acquisition. Researchers have emphasized that

transferability of learning is of prime importance in training. Two types of transfers are discussed in the literature: *near transfer*, in which the testing situation is similar to the training situation, and *far transfer*, in which the testing situation is very different from the training situation [31]. In this study, we focused on measuring near transfer. For example, we trained users regarding revision to their account information from Amazon and tested them on email from Paypal regarding reactivation of their account.

Hypothesis 3: Users transfer more knowledge about how to avoid phishing attacks when trained with embedded training than with non-embedded training.

3.4 Cognitive Reflection

User studies examining phishing or phishing-related interventions are often agnostic to individual user characteristics (sex, age, education level and hours using computer) or have not found significant relationships between features such as age or gender and phishing-related behavior [6, 7, 15, 23]. This may be the product of one or more of the following factors: (1) individual differences (sex, age, education level and hours using computer) are not actually relevant to phishing-related behavior; (2) the sample sizes used for these studies are too small to detect any significant relationships; and (3) truly discriminating characteristics have not yet been tested. In this study, we test previously studied demographic characteristics again, but also investigate whether an individual's propensity for *cognitive reflection* is related to the ability to avoid falling for phishing attacks.

People vary along many dimensions and these variations often result in differences in behavior and decision-making. Frederick suggests that individuals who are more cognitively reflective differ from those who are less reflective [10]. He presents the Cognitive Reflection Test (CRT) consisting of three questions whose correct solutions require the suppression of "impulsivity." In his study, Frederick tested the CRT among approximately 3500 individuals at various universities and in several web-based studies. Although his three-question CRT does correlate highly with other measures of achievement and intelligence such as the Scholastic Aptitude Test (SAT) and the Wonderlic Personnel Test (WPT), Frederick argues that the CRT more accurately measures "cognitive reflection" or "the ability or disposition to resist reporting the response that first comes to mind." He found that higher CRT scores correlate with more risk-taking and lower discount rates. Conversely, those who are less cognitively reflective are more likely to choose certain gains over higher expected values and choose lower amounts immediately over larger rewards later.

The three questions included in the CRT are:

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? ____cents[1]

2. If it takes five machines 5 minutes to make five widgets, how long would it take 100 machines to make 100 widgets? _____minutes[2]

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? ____days[3]

With Cognitive Reflection as our measure of individual variation, we propose two hypotheses about the differential phishing-related behavior users. The first hypothesis draws from the idea that high CRT scores are associated with less impulsive behavior. This hypothesis suggests that individuals with high CRT scores will have a more thorough deliberation process for emails for which they have a mental model. We hypothesize the following:

Hypothesis 4: Users with higher scores on the Cognitive Reflection Test (CRT) will be less likely than users with lower scores to click on "phishing emails" from companies with which they have an account.

On the other hand, the emails ostensibly sent from companies with which a user does not have an account (no-account) are not part of the user's mental model. In this situation, we predict that those with lower CRT scores will be less likely to deviate from the rules and thus not click the links in the no-account emails. On the other hand, we hypothesize that those with a higher CRT score, whom we expect to be greater risk-takers, will explore the no-account emails because of curiosity:

Hypothesis 5: Confronted with a novel situation, those with higher scores on the CRT will be more likely than users with lower scores to click on the links in the phishing emails from companies with which they have no account.

4. EVALUATION

In this section we present the design of the study that we conducted to test the five hypotheses introduced in Section 3. We present our approach to participant recruitment, participant demographics, and study methodology.

4.1 Participant Recruitment and Demographics

We recruited participants by posting fliers in and around our university campus advertising an "email management study." We asked all respondents to complete an online screening survey. We selected people who did not know what phishing was, and who had never taken part in any of our previous studies. Our screening survey included questions like "What does the term 'cookie' mean?" and "Approximately how many times have you used online banking services in the last 6 months?" so that people were not primed towards the idea that we may be conducting a phishing study.

The screening survey was filled out by 165 people; 73 (44.2%) people qualified for the study. Before administering the actual study, we conducted pilot studies with seven qualified participants. The pilot studies were used to refine our study methodology. Forty-nine of the 73 qualified people completed the actual study. However, the data from some participants was excluded from subsequent analysis because they had not viewed the training intervention. Thus we analyzed data for 42 participants who had been randomly assigned to one of three conditions: an "embedded" condition in which participants were shown the training material when they clicked on links in the simulated phishing emails; "non-embedded" condition in which participants were shown training materials in an email message;

---

[1] The correct answer is 5 cents.

[2] The correct answer is 5 minutes.

[3] The correct answer is 47 days.

Table 1: Demographics of the participants; N = 14 in each condition; value presented in parenthesis is standard deviation.

	Embedded Condition	Non-Embedded condition	Control condition
Gender			
Male	36%	43%	36%
Female	64%	57%	64%
Browser			
IE	50%	50%	64%
FireFox	29%	43%	29%
Others	21%	7%	7%
Average emails per day	16 (16.1)	21 (16.8)	21 (17.2)
Average age	25 years (6.9)	24 years (6.9)	28 years (10.2)
Average CRT score	1.25 (0.91)	1.14 (0.94)	1.3 (1.2)
Average time reading the intervention	97 seconds (32.5)	37 seconds (66.2)	-
Average return period	7.2 days (0.9)	7.5 days (0.8)	6.7 days (0.9)

and "control" condition did not receive any training materials but received an additional email from a friend. Table 1 provides the demographic characteristics of the 42 participants whose data we analyzed.

4.2 Methodology

This study was conducted in two laboratory sessions, separated by at least 7 days (mean = 7.2, s.d = 0.95). Participants came to our laboratory for a study investigating "how people effectively manage and use email." When they arrived to our laboratory for the first session we had them fill out the pre-study questionnaire, which included demographic information along with the CRT questions.

Our study consisted of two think-aloud sessions where the participants played the role of "Bobby Smith" the business administrator for Cognix Inc. We had participants sit at a desk in our laboratory, which we told them was Bobby's office desk. The desk was outfitted with a laptop, pens, note pads, post-it notes, and other office supplies. Figure 2 shows the laboratory setup where we conducted the study. We provided the participants with a printout that included details about the role, including names of people Bobby interacts with (co-workers, family, and friends) and all the organizations where Bobby had an account. We also provided the participants with a printout of the user names and passwords for all of Bobby's accounts: AOL, Amazon, American Express, Bank of America, CitiBank, eBay, Gmail, PayPal, Staples, and Yahoo. We showed each participant Bobby's email inbox and asked them to process the email and react to the email as they would in the real world, keeping in mind the role that they are playing. When participants completed the session 1 of the study, no additional information about phishing or nature of the study was provided to the participants.

When participants came back after approximately seven days for the second session, we told them that they would be role playing

Figure 2: One of the participants playing the role of Bobby Smith. The top highlighted box shows the post-it notes that this participant made notes on and stuck to the bookshelf during the user study. The bottom highlighted box shows the participant taking additional notes on the notepad.

Bobby Smith again, just as they had done in session 1. Once again, we showed them Bobby's email inbox and asked them to process Bobby's email. We asked all participants to complete a post-study survey at the end of session 2 after they completed their email management tasks.

We used a 1.70GHz IBM T42 ThinkPad laptop running Microsoft Windows XP Home Edition to conduct the user studies. The participants used Internet Explorer 6.0 for accessing emails through SquirrelMail [27]. We wrote a Perl script to push emails into the SquirrelMail server; and used this script to setup Bobby's inbox for each participant. We recorded the participants' voices and screen-captured their interactions using Camtasia.

We designed the emails in Bobby's inbox to allow us to measure the immediate effectiveness of our interventions as well as knowledge retention and transfer. In session 1, participants saw 33 emails in Bobby's inbox: a set of 16 before-training emails (the "before" set), a training intervention, and a set of 16 additional emails shown immediately after training (the "immediate" set). In session 2, participants saw another 16 emails (the "delay" set) in Bobby's inbox. We had three sets of 16 emails (A, B, and C) that we used for the before, immediate, and delay sets. Each set consisted of 9 legitimate emails without any links in them from people with whom Bobby interacts (legitimate-no-link), 3 legitimate emails containing links from organizations and people with whom Bobby interacts (legitimate-link), 2 phishing emails from organizations where Bobby has an account (phishing-account), 1 email from a bank with which Bobby does not have an account (phishing-no-account), and 1 spam email. Participants were randomly assigned to see either set A or set C as the before set and the other one as the delay set. All participants saw set B as the immediate set. Table 2 summarizes the contents of email set A. Sets B and C contained the same types of emails with a different combination of senders and subjects.

All participants in the embedded and non-embedded training conditions saw a training intervention from Amazon, a company with which Bobby had an account, with the subject "Revision to your Amazon.com information." Participants in the embedded condition saw the training material shown in Figure 1 when they

Table 2. Arrangement of email in set A. The other sets had similar distribution of emails. LNL = legitimate-no-link LL = legitimate-link, PA = phishing-account, PNA = phishing-no-account.

Email position	Sender information	Email subject line information	LNL	LL	PA	PNA	Spam
1	Josept Dicosta	[cognix] REMINDER: Don't forget to attend the tax session	√				
2	Ni Cheng	RE: Room booking - Sunday - To meet - Let me know	√				
3	PayPal	Reactivate you PayPal account!			√		
4	Brandy Anderson	Booking hotel rooms for visitors	√				
5	Jean Williams	Re: Funny joke (fwd)	√				
6	Eddie Arredondo	Fw: Re: You will want this job					√
7	Brandy Anderson	To check the status of the product on Staples		√			
8	Fiona Jones	Don't forget mom's birthday!	√				
9	Wells Fargo	Update your bank account information!				√	
10	Brandy Anderson	Please check Paypal balance		√			
11	Jean Williams	coffee from starbucks	√				
12	Ni Cheng	RE: Tea powder - Kitchen	√				
13	AOL	IMPORTANT: Please Update Your AOL account			√		
14	Brandy Anderson	New member in our administrative team	√				
15	American Express	Confirmation: Payment Received		√			
16	Jesse	Sorry missed your call - will call you this weekend	√				

clicked on the link in the email, while the non-embedded condition received the training message in the email itself. Participants in the control condition did not receive any training material but they received an email from a friend.

All the phishing, spam, and legitimate-with-link emails that we used for this study were based on actual emails that we had collected from members of our research group. We designed the legitimate-no-link emails to resemble the emails that one of our business administrators typically receives. We created exact replicas of the phishing websites on our local machine by running Apache and modifying the host files in Windows so that IE would display the URL of the actual phishing websites. All replicated phishing websites were completely functional and allowed people to submit information. These phishing websites were only accessible from the laboratory machine used for the user studies. Users were taken to these phishing websites when they clicked on links in the phishing-account and phishing-no-account emails.

5. RESULTS

In this section we present the results of the user study we conducted to test the five hypotheses introduced in Section 3. We consider someone to have fallen for a phishing attack if they click on a link in a phishing email, regardless of whether they go on to provide personal information. The conclusions presented in this paper are robust to the selection of a different metric for the evaluation of the correctness of the participants' choices. Specifically, the findings listed in this section persist when we focus on the participants who provided personal information to a

phishing website during the experiment, rather than simply clicking on the links of a spoofed email. (Although not everyone who clicks on a phishing link will go on to provide personal information to a website, in this study people who clicked on phishing links provided information 90% of the time.) We calculated correctness scores as the number of emails containing links that a participant correctly identified as phishing or legitimate. We determined the correctness of the identification based on whether or not the participant clicked on a link in each email.

The results from the study supported hypotheses 1, 2, 3, and 5; and rejected hypothesis 4. We found no correlation between the participants' scores (correctly identifying phishing emails as phishing and legitimate emails as legitimate) and participants' demographics. We found that participants in the embedded condition made better decisions after the training compared to participants in the non-embedded condition. In fact, participants in the non-embedded condition did not perform significantly better after training than those in the control condition (who had received no training). Also, participants in the embedded condition spent significantly more time reading the intervention than participants in the non-embedded condition. We found that participants in the embedded condition retained and transferred more knowledge than participants in non-embedded condition. We found that participants with higher Cognitive Reflection Test (CRT) scores are more likely than users with lower CRT scores to click on the links in the phishing emails from companies with which they have no account. We also found that participants
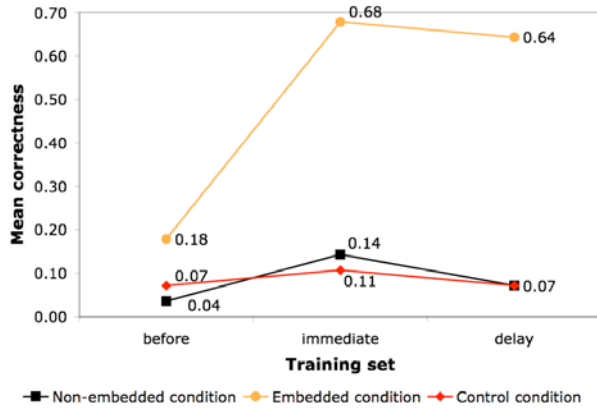
Figure 3: The mean correctness for the phishing-account emails before and immediately after training and after a one-week delay.

Figure 4: The mean correctness for the legitimate-link emails before and immediately after training and after a one-week delay. There was no significant difference among the three conditions.

generally liked the embedded training methodology and the intervention design (comic strip) that we used for the study.

5.1 Participant scores and behavior

We determined the number of correct decisions that participants made about the six emails in each set that contained links and the one spam email to calculate a score between 0 and 7 for each participant on each email set. We counted a decision about a legitimate email as correct if the participant clicked on the link and performed the requested action. We counted a decision about a phishing email as correct if the participant did not click on the link in the email. We counted a decision about a spam email as correct if the participant did not open the email. We also calculated the percentage correct for each participant and each type of email in each set. We present the average percentage correct for each email in Table 3 in Appendix.

Before the training, we found no significant difference (t = 1.48, p-value = 0.17) in scores for the phishing-account messages in email sets A and C, indicating they were of similar difficulty. Within each group we found no significant difference between the scores for the two phishing-account emails that the participants received (proportion test: A group, p-value = 0.37, and C group, p-value = 0.32). This shows that the phishing-account emails presented in a group did not differ significantly.

Among the seven participants who were excluded from our analysis because they did not look at the training materials, three were in the non-embedded condition and four were in the embedded condition. Among the four in the embedded condition, two participants did not open the training email and two of them did not click on the link in the email. The three participants in the non-embedded condition did not open the email. The total correctness score for participants who did not look at the intervention was 6.33 for the embedded condition and 6.25 for the non-embedded condition. We found a significant difference between the scores for people who saw the training and people who did not see the training material. The responses of these seven participants were not included in the analysis discussed in this paper.

We found no significant correlation between phishing susceptibility and the demographic information that we collected. For instance, there was no significant correlation between
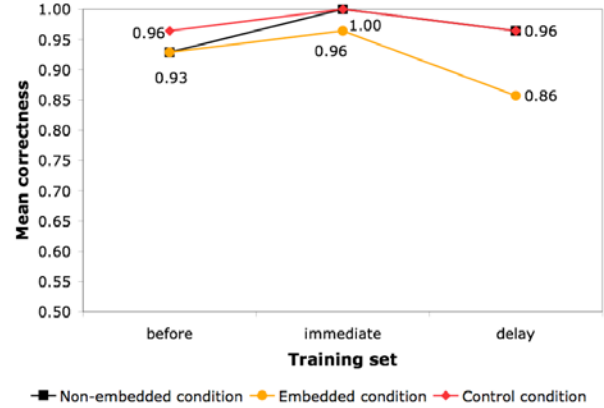
participant's age and total scores (Pearson coefficient r = 0.30, p-value = 0.13). There was no significant correlation between emails received per week (excluding unsolicited) and total scores (Pearson coefficient r = 0.02, p-value = 0.92). There was no significant correlation between shopping online in the last six months and score (Pearson coefficient r = -0.12, p-value = 0.56). There was no significant correlation between hours of Internet usage per week and score (Pearson coefficient r = 0.24, p-value = 0.22). There was also no significant difference in scores between males and females (t = -1.1, p-value = 0.29). The mean score of males was 4.27 (s.d = 1.19, var = 1.42) and the mean score for females was 4.71 (s.d = 0.69, var = 0.47). We also observed no significant difference between the non-embedded condition and the control condition (details in Figure 3 and Figure 4).

5.2 Learning

In this section we assess how much users learned as a result of the interventions.

5.2.1 User performance

To test hypothesis 1, the effectiveness of the training was evaluated using the percentage correct score of participants in each condition for phishing and legitimate-link emails before and after the training.

Participants in the embedded and non-embedded conditions did not perform significantly different in correctly identifying phishing-account emails before the training (two sample t-test: df = 26, p-value = 0.19). However those in the embedded condition performed significantly better than those in the non-embedded condition immediately after training (two sample t-test: df = 26, p-value < 0.01), as shown in Figure 3. Those in the embedded condition improved their performance significantly immediately after the training (paired t-test: t = -3.61, df = 13, p-value < 0.01), while those in the non-embedded condition did not (paired t-test: t = -1.15, df = 13, p-value = 0.27). There was no significant difference between the control condition and the non-embedded condition both before and after the training.

Participants in the embedded and non-embedded conditions did not perform significantly differently in correctly identifying legitimate-link emails before or after the training, as shown in Figure 4. There was no significant difference for mean correctness between before and immediately after the training in embedded

(paired t-test: t = -1, df = 13, p-value = 0.34) and non-embedded condition (paired t-test: t = -1.47, df = 13, p-value = 0.17). Similarly, there was no significant difference for mean correctness between non-embedded and the control condition.

Our results support Hypothesis 1, demonstrating that embedded training increases users ability to detect phishing-account emails while non-embedded training does not. No form of training had significant impact on user's ability to recognize legitimate emails.

5.2.2 Time spent in reading the intervention
One approximate measure for how closely people read the training materials is the time spent looking the materials. Learning science suggests that users exposed to training materials for more time may learn more [31]. We measured the time participants spent on the training materials in each condition. There was significant difference (two sample t-test: t = -3, df = 26, p-value < 0.01) between the embedded condition (min = 21 seconds, max = 240 seconds, avg = 97 seconds) and the non-embedded condition (min = 2 seconds, max = 100 seconds, avg. = 37 seconds). This shows that participants in the embedded condition spent significantly more time reading the training material compared to participants in the non-embedded condition. We also found significant correlation between the time spent in reading the training material and the total scores immediately after the training (Pearson coefficient r = 0.6, p-value < 0.01) and also after the delayed time period (Pearson coefficient r = 0.44, p-value = 0.02).

5.3 Retention and transfer
In order to measure retention and transfer, we asked participants to come back for a second part of the study. We requested that they come back exactly 7 days after part 1. However, not all of the participants came back in exactly seven days. The participants from the non-embedded condition came an average of 7.5 days apart (min = 6, max = 9, s.d = 0.94, var = 0.88). Embedded condition participants on average came back after 7.2 days (min = 6, max = 9, s.d = 0.80, var = 0.64). Control condition participants on average came back after 7.1 days (min = 6, max = 9, s.d = 0.7, var = 0.5). There was no significant difference for the days apart between the three conditions.

5.3.1 Overall performance after a delay
In order to measure overall user performance after the one-week delay, we compared correctness percents for phishing-account and legitimate-link emails before, immediately after training, and after one-week delay. Participants in the embedded condition performed significantly better then those in the non-embedded condition even after one-week delay (two sample t-test: df = 26, p-value < 0.01), as shown in Figure 3. Participants in the embedded condition improved their performance significantly after the delay compared to before the training (paired t-test: t = -2.51, df = 13, p-value = 0.02), while participants in the non-embedded group did not improve (paired t-test: t = -0.43, df = 13, p-value = 0.67). In both the conditions there was no significant difference between immediately after the training and after a delay of one-week. Participants in the control condition did not perform significantly better after the delay compared to immediately after the training. In fact, the mean correctness score after the delay was exactly the same as before the training.

Participants in the embedded, non-embedded, and control conditions did not perform significantly differently in correctly identifying legitimate-link emails after the delay as shown in

Figure 4. There was no significant difference for mean correctness between before the training and after the delay in all the three conditions.

These results suggest that users were able to identify phishing and legitimate emails correctly better in the embedded condition than in the non-embedded and the control condition even after a delay of one-week.

5.3.2 Retention
The intervention email appeared to be from Amazon with the subject "Revision to your Amazon.com information." This email requests the user to update the personal information for their Amazon account. To measure retention (similar type of phishing email after a delay) we used an email from Citibank requesting users to update their personal information for the account. There was a significant difference between the non-embedded and the embedded training condition for identifying correctly the email from Citibank as phishing email (two sample t-test: df = 26, p-value < 0.01). There was also significant difference between the embedded and the control condition. This result lends support to Hypothesis 2. Only 7% of the participants identified the email correctly in the non-embedded and the control condition, while 64% of the participants identified the email correctly in the embedded condition. One of the participants in the embedded condition mentioned that "I remember reading last time that thing [training material] said not click and give personal information."

5.3.3 Transfer
To measure the knowledge transfer (different type of phishing email after a delay) we used an email (phishing-account type) that asked participants to reactivate their eBay account. We found significant differences between the non-embedded and the embedded training conditions in correctly identifying the eBay email as a phishing attack (two sample t-test: df = 26, p-value < 0.01). This result lends support to Hypothesis 3. Only 7% of the participants identified the email correctly in the non-embedded and the control condition, while 64% of the participants identified the email correctly in the embedded condition. One of the participants in the embedded condition mentioned that "PhishGuru said not to click on links and give personal information, so will not do it, I will delete this email."

5.4 Cognitive Reflection
As mentioned earlier, we included Frederick's three-question Cognitive Reflection Test (CRT) as part of our pre-screening survey. The raw CRT score ranged from 0 to 3, with "0" indicating that the subject did not answer any of the three questions correctly and "3" indicating that the subject answered all three correctly. The mean CRT score was 1.23 and s.d. = 0.95. We dichotomized the CRT score by converting CRT scores of 0-1 to "low CRT group" and 2-3 to "high CRT group." We had 24 subjects in the low CRT group and 18 in the high CRT group. There was no significant difference between the three conditions regarding the means. We also found no significant correlation between the age of the participants and the CRT score (Pearson coefficient r = -0.16, p-value = 0.32).

We tested our hypotheses by comparing the proportion of individuals in the two CRT groups (high and low) who clicked the phishing-account and phishing-no-account emails prior to training using a test of two proportions. For Hypothesis 4, we predicted

that the high CRT group had a lower probability of clicking on the phishing e-mail from the company with whom they have an account. Using a test of 2-proportions, we found a difference exists in the predicted direction; however, our statistical analysis suggests that this difference between the proportion clicking on the phishing email for the low CRT group (0.92) and the high CRT group (0.72) is not significant (proportion test: p-value = 0.10). This result rejects Hypothesis 4.

In the case of Hypothesis 5, we expected that subjects who had higher CRT scores would be more likely to click on the phishing-no-account emails prior to training. We conjectured that because higher CRT scores correlate with more risk-taking, the high CRT subjects would be more likely to click on the e-mails that were unexpected, given the Bobby Smith storyline. In our sample, the high CRT group had a higher probability of clicking on the phishing-no-account e-mails than those in the high CRT group, 0.39 versus 0.04, respectively. A test of 2-proportions suggests that the difference in proportions was significant (p-value < 0.01). These results indicate that those with high CRT scores are more likely to click on phishing-no-account e-mails than those with low CRT scores. This result lends support to Hypothesis 5. It does not mean that those who are more "cognitively reflective" are more likely to fall for phishing attacks, but may suggest that they are more inclined to "play with fire." In a novel situation, they may be more inclined to experiment and have a higher level of curiosity about unknown e-mails than those with lower CRT scores. However, this may or may not suggest that those with high CRT scores are more likely to be "burned." In a real situation, although they may be curious about the e-mail, its content, and the website it links to, they may not necessarily enter their personal information into a website they do not trust. Nevertheless, clicking on the email may expose the individual to other types of security threats such as viruses.

5.5 Observations

Participants in both conditions identified spam emails correctly most of the time, meaning that users did not even open the email. Almost all (93%) participants identified the spam email correctly before training in both conditions. One of the participants who opened the spam email was curious about it (subject of the email: "Fw: Re: You will want this Job"). Another participant said "Oh, it is offering me a job, might be interesting, let me see it." There was no significant difference within the conditions for before, immediate and a delay of training. There was also no significant difference between the conditions in any of the states.

There was a significant difference in correctness among participants for the phishing emails from organizations that they have an account with (phishing-account) and for the emails that they do not have an account with (phishing-no-account). There was a significant difference between the phishing-account and phishing-no-account emails within the conditions before the training. One of the common reasons mentioned by the participants for not opening or for deleting the phishing-no-account emails is "I don't have an account with this organization." In particular, one of the participants mentioned, "I don't have account with Barclays, how did they get my email address, and why are they sending emails asking me to update my information."

We observed that participants in the embedded condition were motivated to read the training material longer than the non-embedded condition. One participant mentioned, "I was more

motivated to read the training materials since it was presented after me falling for the attack." This quote succinctly captures the motivation behind the embedded training methodology, which makes training part of users' primary task. Another participant in the embedded condition mentioned, "Thank you PhishGuru, I will remember that [the 5 instructions given in the training material]." In general, participants who spent time on reading the training material liked the design. One participant who was not aware that URLs could be misleading looked at the arrow pointing to the first "n" in "amazon.com" (Figure 1) and said, "That is scary, I will be careful in the future. That [instruction] is good to know." The non-embedded condition does not create the motivation as in the embedded condition. This can be seen in one of the participant's comment "This [image in the email] looks like some spam." Another participant mentioned "I don't know why Amazon would send me such [intervention] in the email."

6. DISCUSSION

The results from the study supported hypotheses 1, 2, 3 and 5, and rejected hypothesis 4 mentioned in Section 3. Results from this study contradict the conventional wisdom that it is hard to train users about security. Our results are consistent with learning literature findings that users can be trained if the methodology is systematically designed and applies learning science principles.

Our results strongly suggest that sending instructional materials through email (non-embedded) does not motivate users to spend time on the instructions. We believe this is because people are unclear as to why they are receiving such emails and so delete the emails with the instructions. Our results also suggest that users are motivated to learn when the training materials are presented after users fall for the phishing emails (when users click on the link in the email). We believe this is because the embedded methodology directly applies the learning-by-doing and immediate feedback principles.

Our results suggest that users can retain and transfer knowledge if they are motivated to read the training materials. Our results indicate that after seven days participants in the embedded condition retained the knowledge that they gained better than the participants in the non-embedded condition. This may suggest that creating motivation by making users fall for phishing emails influences their retention of knowledge. We also found that participants in the embedded condition were able to transfer their knowledge to a different situation than the trained situation better than the participants in the non-embedded condition. This suggests that if users were frequently trained on phishing emails, they should be able to identify other types of phishing emails.

The results from the post-study discussion with participants showed that almost all participants liked the comic script intervention design that we used for this study. We attribute this to the learning science principles (learning-by-doing, immediate feedback, contiguity, personalization, and story-based agent) that we applied for creating the design.

Our analysis found users with high and low CRT scores were equally likely to click on the links in the phishing emails from organizations that they have an account with. Our analysis also found that participants with high CRT scores are more likely to click on phishing emails when they are from an unknown source. This result may indicate that training the high CRT score group not to click on links from unknown sources may be appropriate.

7. CONCLUSIONS AND FUTURE WORK

In this paper we showed that: (a) users learned more effectively when the training materials were presented after they fell for the phishing attack (embedded) than when the training materials were sent by email (non-embedded); (b) users retained more knowledge when trained with embedded training than when trained with non-embedded training; (c) users transferred more knowledge about how to avoid phishing attacks when trained with embedded training than when trained with non-embedded training; (d) users with high and low Cognitive Reflection Test (CRT) scores had equal chances to click on the links in the phishing emails from organizations that they have an account with (phishing-account emails); and (e) users with high CRT scores were more likely than users with lower scores to click on links in emails from an organization that they do not have an account with (phishing-no-account emails), perhaps due to their curiosity.

We are currently working on designing instructional materials for other cues and strategies that users can be trained on. We are applying the learning science principles discussed in this paper for designing future materials also. We also plan to conduct a field trial of this system. In such real-world setup, we plan on making the training system adaptive to the users' knowledge and skills.

8. ACKNOWLEDGMENTS

9. REFERENCES

1. Anandpara, V., Dingman, A., Jakobsson, M., Liu, D., and Roinestad, H. Phishing IQ tests measure fear, not ability. Usable Security (USEC'07) (2007). http://usablesecurity.org/papers/anandpara.pdf.

2. Anderson, J. R. Rules of the Mind. Lawrence Erlbaum Associates, Inc., 1993.

3. Anderson, J. R., and Simon, H. A. Situated learning and education. Educational Researcher 25 (1996), 5–11.

4. Anton, A. I., Earp, E. A. J. B., Bolchini, D., He, Q., Jensen, C., and Stufflebeam, W. The Lack of Clarity in Financial Privacy Policies and the Need for Standardization. IEEE Security and Privacy 2(2) (2004), pp. 36–45. Retrieved Dec 20, 2004, http://www.theprivacyplace.org/papers/glb_secPriv_tr.pdf.

5. Clark, R. C. and E. M. Richard. 2002. *E-Learning and the science of instruction: proven guidelines for consumers and designers of multimedia learning.* Pfeiffer, San Francisco, USA.

6. Dhamija, R., Tygar, J. D., and Hearst, M. 2006. Why phishing works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada, April 22 - 27, 2006). R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson, Eds. CHI '06. ACM Press, New York, NY, 581-590. DOI= http://doi.acm.org/10.1145/1124772.1124861.

7. Downs, J. S., Holbrook, M. B., and Cranor, L. F. 2006. Decision strategies and susceptibility to phishing. In *Proceedings of the Second Symposium on Usable Privacy and Security* (Pittsburgh, Pennsylvania, July 12 - 14, 2006). SOUPS '06, vol. 149. ACM Press, New York, NY, 79-90. DOI= http://doi.acm.org/10.1145/1143120.1143131.

8. eBay Toolbar. Retrieved December 30, 2006. http://pages.ebay.com/ebay_toolbar/

9. Fette, I., N. Sadeh and A. Tomasic. Learning to Detect Phishing Emails. June 2006. ISRI Technical report, CMU-ISRI-06-112 (To be presented at WWW 2007). http://reports-archive.adm.cs.cmu.edu/anon/isri2006/CMU-ISRI-06-112.pdf.

10. Frederick, S. Cognitive reflection and decision making. *Journal of Economic Perspectives* 19, 4 (2005), 25–42.

11. Keinan, G. Decision making under stress: scanning of alternatives under controllable and uncontrollable threats. *Journal of personality and social psychology* 52, 3 (1987), 639–644.

12. Kirkley, J. R., and et al. Problem-based embedded training: An instructional methodology for embedded training using mixed and virtual reality technologies. In Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) (2003). http://www.iforces.org/downloads/problem-based.pdf.

13. Klein, G. *Sources of power: How people make decisions?* The MIT Press Cambridge, Massachusetts The MIT Press, Cambridge, Massachusetts, London, England, February 1999.

14. Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., and Hong, J. Teaching johnny not to fall for phish. Tech. rep., Cranegie Mellon University, 2007. http://www.cylab.cmu.edu/files/cmucylab07003.pdf.

15. Kumaraguru, P., Y. Rhee, A. Acquisti, L. Cranor, J. Hong, and E. Nunge. In *Proceedings of CHI 2007*. Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System.

16. Mayer, R.E. *Multimedia Learning.* 2001. New York Cambridge University Press.

17. Mayer, R. E., and Anderson, R. B. The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology* 84, 4 (December 1992), 444–452.

18. Merrienboer, J. V., de croock, M., and Jelsma, O. The transfer paradox : Effects of contextual interference on retention andtransfer performance of a complex cognitive skill. Perceptual and motor skills 84 (1997), 784–786.

19. Moreno, R., Mayer, R. E., Spires, H. A., and Lester, J. C. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction* 19, 2 (2001), 177–213.

20. Robila, S. A., J. James and W. Ragucci. 2006. Don't be a phish: steps in user education. ITICSE '06: *Proceedings of the 11th annual SIGCSE conference on Innovation and technology in computer science education*. pp 237-241. New York, NY, USA.

21. Rubin, D. C., and Wenzel, A. E. One hundred years of forgetting : A quantitative description of retention. *Psychological Review* 103, 4 (1996), 734–760.

22. Schmidt, R. A., and Bjork, R. A. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science* 3, 4 (July 1992), 207–217.

23. Sheng, S., B. Magnien, P. Kumaraguru, A. Acquisti, L. Cranor, J. Hong, and E. Nunge. Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. To appear in *Symposium on Usable Privacy and Security* 2007.

24. SpamAssasin. Retrieved September 10, 2006. http://spamassassin.apache.org/

25. SpoofGuard. Retrieved September 10, 2006, http://crypto.stanford.edu/SpoofGuard/

26. SpoofStick. Retrieved September 10, 2006. http://www.spoofstick.com/

27. SquirrelMail. Retrieved September 10, 2006. http://www.squirrelmail.org/

28. Tversky, A., and Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science 185*, 4157 (1974), 1124–1131.

29. Tversky, A., and Shafir, E. The disjunction effect in choice under uncertainty. *American Psychological Society 3*, 5 (September 1992), 305 – 309.

30. Whitten, A and J. D. Tygar. 1999. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. Proceedings of the 8th USENIX Security Symposium. http://www.cs.berkeley.edu/~tygar/papers/Why_Johnny_Cant _Encrypt/USENIX.pdf.

31. Whitten, W. B., and Bjork, R. A. Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior* 16, 4 (August 1977), 465–478.

10. APPENDIX

Table 3: The percentage of users who correctly identified each email. A user who clicked on a link in a legitimate email or refrained from clicking on a link in a phishing email or did not open the spam email was deemed to have made a correct identification. Columns with (-) indicate that the email set was not used in that state.

Email type	Non-embedded condition			Embedded condition			Control condition		
	Before set	Immediate set	Delay set	Before set	Immediate set	Delay set	Before set	Immediate set	Delay set
Email set A									
Legitimate-link-1	1	-	0.86	0.83	-	0.86	0.86	-	1
Phishing-account-1	0.14	-	0.14	0.17	-	0.63	0.26	-	0
Spam	0.86	-	0.86	1	-	0.75	1	-	1
Legitimate-link-2	0.86	-	1	1	-	0.63	0.86	-	1
Phishing-no-account	0.71	-	0.71	1	-	1	1	-	0.43
Phishing-account-2	0	-	0.14	0	-	0.63	0	-	0.14
Legitimate-link-3	0.86	-	1	1	-	1	1	-	1
Email set B									
Legitimate-link-1	-	1	-	-	1	-	-	1	-
Phishing-account-1	-	0.14	-	-	0.67	-	-	0.14	-
Spam	-	1	-	-	0.88	-	-	1	-
Legitimate-link-2	-	1	-	-	0.94	-	-	1	-
Phishing-no-account	-	0.79	-	-	0.79	-	-	0.93	-
Phishing-account-2	-	0.14	-	-	0.73	-	-	.07	-
Legitimate-link-3	-	0.57	-	-	0.48	-	-	0.86	-
Email set C									
Legitimate-link-1	0.71	-	0.71	0.5	-	0.67	1	-	1
Phishing-account-1	0	-	0	0.25	-	0.67	0.14	-	0.14
Spam	1	-	1	0.88	-	0.83	1	-	1
Legitimate-link-2	1	-	1	0.75	-	0.83	1	-	0.86
Phishing-no-account	0.86	-	0.86	0.63	-	1	0.86	-	0.71
Phishing-account-2	0	-	0	0.25	-	0.67	0	-	0
Legitimate-link-3	1	-	1	1	-	1	1	-	1
Average for each type of email across all email sets									
Legitimate-link	0.90	0.86	0.95	0.84	0.81	0.83	0.95	0.95	0.98
Phishing-account	0.03	0.11	0.07	0.17	0.70	0.65	0.11	0.09	0.07
Phishing-no-account	0.79	0.79	0.79	0.81	0.79	1	0.93	0.93	0.57
Spam	0.93	1	0.93	0.93	0.88	0.79	1	1	1