Chapter 5

Anti-Phishing Phil: A Case study in User education

> This chapter is joint work with Alessandro Acquisti, Lorrie Cranor, Jason Hong, and Ponnurangam Kumaraguru. An earlier version of the content in this chapter was published at SOUPS 2007 [127] .

5.1 Introduction

Phishing is a kind of attack in which criminals use spoofed emails and fraudulent web sites to trick people into giving up personal information. Victims perceive these emails as associated with a trusted brand, while in reality they are the work of con artists interested in identity theft [57]. These increasingly sophisticated attacks not only spoof email and web sites, but they can also spoof parts of a user's web browser [55].

Phishing is part of a larger class of attacks known as *semantic attacks*. Rather than taking advantage of system vulnerabilities, semantic attacks take advantage of the way humans interact with computers or interpret messages [123], exploiting differences between the system model and the user model [139]. In the phishing case, attacks exploit the fact that users tend to trust email messages and web sites based on superficial cues that actually provide little or no meaningful trust information [26, 55].

Automated systems can be used to identify some fraudulent email and web sites. However, these systems are not completely accurate in detecting phishing attacks. In a recent study, only one of the ten anti-phishing tools tested was able to correctly identify over 90% of phishing web sites, and that tool also incorrectly identified 42% of legitimate web sites as fraudulent [147]. It

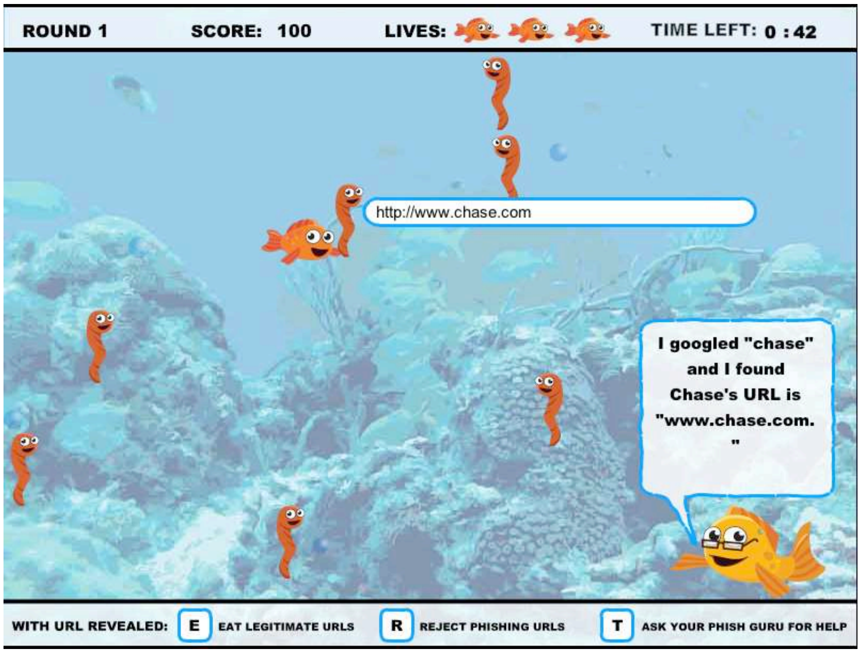


Figure 5.1 Anti-Phishing Phil game screen. Phil, the small fish near the top of the screen, is asked to examine the URL next to the worm he is about to eat and determine whether it is associated with a legitimate web site or a phishing site. Phils father (lower right corner) offers some advice. The game is available at: `http://cups.cs.cmu.edu/antiphishing_phil/`

is also unlikely that any system will ever be completely accurate in detecting phishing attacks, especially when detection requires knowledge of contextual information. While it makes sense to use automated detection systems as one line of defense against semantic attacks, our philosophy is that there will still remain many kinds of trust decisions that users must make on their own, usually with limited or no assistance. The goal of our research is not to make trust decisions for users, but rather to develop a complementary approach to *support*users so that they can make better trust decisions. More specifically, one goal of our research is to find effective ways to train people to identify and avoid phishing web sites.

In this paper we present the design, implementation, and evaluation of Anti-Phishing Phil, a game we developed to teach people how to protect themselves from phishing attacks. Anti-Phishing Phil teaches people how to identify phishing URLs, where to look for cues in web browsers, and how to use search engines to find legitimate sites. In Section 2, we present background information and related work on why people fall for phishing, and approaches to protecting them. In Section 3, we describe the design of Anti-Phishing Phil, and present the ways in which we applied learning principles in designing the game. In Section 4, we present the methodology we used to evaluate the game. In Section 5, we present the results of our evaluation, which shows that the game is more effective than a tutorial we created or existing online training materials at teaching people to identify phishing web sites accurately. We discuss the effect of anti-phishing training in Section 6. Finally, we present our conclusions in Section 7.

5.2 Background and Related Work

In this section, we present background on anti-phishing research, why people fall for phishing, and approaches to protecting people from falling for phishing attacks. Previous work on phishing falls into three categories: studies to understand why people fall for phishing attacks, tools to protect people from such attacks, and methods for training people not to fall for phishing attacks.

5.2.1 Why people fall for phishing

Downs et al have described the results of an interview and role-playing study aimed at understanding why people fall for phishing emails and what cues they look for to avoid such attacks. There were two key findings in their work. First, while some people are aware of phishing, they do not link that awareness to their own vulnerability or to strategies for identifying phishing attacks. Second, while people can protect themselves from familiar risks, people tend to have difficulties generalizing what they know to unfamiliar risks [26].

Dhamija et al showed twenty-two participants twenty web sites and asked them to determine which were fraudulent. Participants made mistakes on the test set 40% of the time. The authors noted that 23% of their participants ignored all cues in the browser address bar and status bar as well as all security indicators [23]. This study did not present users with the email messages that might lead users to visit the web sites presented, so it provides no data on whether users pay attention to, or how they interpret, email cues.

Wu et al. studied three simulated anti-phishing toolbars to determine how effective they were at preventing users from visiting web sites the toolbars had determined to be fraudulent. They found that many study participants ignored the passive toolbar security indicators and instead used the site’s content to decide whether or not it was a scam. In some cases participants did not notice warning signals, and in other cases they noticed them but assumed the warnings were invalid. In a follow-up study, the authors tested anti-phishing toolbars that produced pop-up warnings that blocked access to fraudulent web sites until overridden by the user. These pop-up warnings reduced the rate at which users fell for fraudulent sites, but did not completely prevent all users from falling for these sites. The authors concluded that Internet users are not very good at interpreting security warnings and are unfamiliar with common phishing attacks, and recommended educating users about online safety practices [140].

Our work builds on these previous studies. We incorporated many of the lessons learned from this past work into our game. For example, we teach people not to trust the content of the web page but examine the URL instead. Our evaluation methodology is also closely based on Dhamija et al.’s work [23].

5.2.2 Tools to protect people from phishing

Anti-phishing services are now provided by Internet service providers, built into mail servers and clients, and available as web browser toolbars. However, these services and tools do not effectively protect against all phishing attacks, as attackers and tool developers are engaged in a continuous arms race [147]. Furthermore, Internet users who are unaware of the phishing threat will be unlikely to install and use an anti-phishing tool, and may ignore warnings from anti-phishing tools provided by their ISPs or built into their web browsers. Even users who understand anti-phishing warnings may ignore them [140]. Where possible, anti-phishing tools should be applied, but—as noted in the introduction—there will always be cases where people have to make trust decisions on their own.

Other research has focused on the development of tools to help users determine when they are interacting with a trusted site. Ye et al. [145] and Dhamija and Tygar [22] have developed prototype “trusted paths” for the Mozilla web browser that are designed to assist users in verifying that their browser has made a secure connection to a trusted site. Herzberg and Gbara have developed

TrustBar, a browser add-on that uses logos and warnings to help users distinguish trusted and untrusted web sites [51]. Other tools, such as PassPet and WebWallet, try to engage users by requiring them to interact actively with the tool before giving out sensitive information [139], [141], [140]. However, even these solutions ultimately rely on the user's ability to make the right decision. In addition, these approaches require either end-users, web servers, or both to install special software. In contrast, our training method only relies on teaching people what cues to look for in existing web browsers.

5.2.3 Anti-phishing education

Despite claims by security and usability experts that user education about security does not work [31], there is evidence that well designed user security education can be effective [68]. Web-based training materials, contextual training, and embedded training have all been shown to improve users' ability to avoid phishing attacks.

A number of organizations have developed online training materials to educate users about phishing [28], [32]. In a previous study, we tested the effectiveness of some of these online materials and found that, while these materials could be improved, they are surprisingly effective when users actually read them [70].

Several studies have adopted a *contextual training* approach in which users are sent simulated phishing emails by the experimenters to test users' vulnerability to phishing attacks. At the end of the study, users are given materials that inform them about phishing attacks. This approach has been used in studies involving Indiana University students [53], West Point cadets [33], and New York State employees [104]. In the New York State study, employees who were sent the simulated phishing emails and follow-up notification were better able to avoid subsequent phishing attacks than those who were given a pamphlet containing information on how to combat phishing.

A related approach, called *embedded training*, teaches users about phishing during their regular use of email. In a previous laboratory experiment to evaluate our prototype embedded training system, we asked our participants to role play and respond to the messages in an email inbox that included two training emails designed to look like phishing emails. If a participant clicked on a link in a training email, we immediately presented an intervention designed to train them not to fall for phishing attacks. We created several intervention designs based on learning sciences, and found that our interventions were more effective than standard security notices that companies email to their customers [68].

We designed our anti-phishing game to complement the embedded training approach, which trains people while they are performing their primary task (checking email). If users are interested in devoting some additional time to learning more about phishing, they can play the Anti-Phishing Phil game. The embedded training approach trains users to identify phishing emails, while the game teaches users to identify phishing web sites. The game emphasizes that phishing web sites often can be identified by looking at their URLs, and teaches users about the various parts of a URL. This training may also help users analyze URLs in suspicious email messages.

5.3 Design of Anti-phishing Phil

In this section we present: the objectives of the game; learning science principles that we applied in designing the game; the story, mechanics, and technology of the game; and results from some of the pilot studies that we conducted, as we iterated on the game design.

We used an iterative design process to develop the game. Our early iterations made use of paper and Flash prototypes to explore various design alternatives. After a great deal of play-testing and feedback from our research group, both on the content of the game (what to teach) and the game design itself (presentation), we developed a working prototype that we tested with actual users. We then iterated on the design several more times based on user feedback and behavior, focusing on improving the game mechanics and messages. Finally, we created a more polished look and feel using attractive images and enticing sounds.

5.3.1 Game Design Principles

In this section, we present the objectives for the game and the learning science principles that we applied in implementing these objectives.

Our objective in developing the anti-phishing game was to teach users three things: (1) how to identify phishing URLs, (2) where to look for cues for trustworthy or untrustworthy sites in web browsers, and (3) how to use search engines to find legitimate sites. We believe that search engines can be an effective tool in identifying phishing web sites. For example, users can search for a brand name in a search engine and see whether the link that appears in the top search results is the same as a potentially suspicious link received in an email. By far, the top search engine results are legitimate web sites [148].

To achieve the above-mentioned objectives, we applied several learning science principles to the game design. Learning sciences theory suggests that training will be effective if the training methodology is goal-oriented, challenging, contextual, and interactive [110]. In goal-oriented training, learners have a specific goal to achieve and in the process of achieving the goal they are challenged and trained. Training is most effective if the materials are presented in a context users can relate to, and if the materials are presented in an interactive form. There also exists a large body of literature on the effectiveness of games for interactively teaching *conceptual* and *procedural* knowledge [44]. Conceptual knowledge is knowledge about concepts or relationships that can be expressed as propositions (e.g., URLs have a protocol part and a domain name part). In contrast, procedural knowledge (also referred as declarative knowledge) is the step-by-step knowledge that one uses to solve a given problem (e.g., check the URL in the address bar, and if it contains an IP addresses, you are likely visiting a phishing site) [3]. The Anti-Phishing Phil game conveys both conceptual and procedural knowledge. Research in learning science has established that interactive environments, in particular games, are one of the most effective training methods and are highly motivational for users, especially when they adhere to design principles for educational games [44], [110], [113]. We applied three learning science principles to the design of the Anti-Phishing Phil game: reflection, story-based agent, and conceptual–procedural.

Reflection principle. Reflection is the process by which learners are made to stop and think about what they are learning. Studies have shown that learning increases if educational games include opportunities for learners to reflect on the new knowledge they have learned [18]. This principle is employed in our anti-phishing game by displaying, at the end of each round, a list of web sites that appeared in that round and whether the user correctly or incorrectly identified each one (as shown in Figure 2). This helps users reflect on the knowledge gained from the round they just completed.

Story-based agent environment principle. Agents are characters that help in guiding learners 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 as part of story-based content enhances learning. We applied this principle in the game by having the user control a young fish named Phil, who has to learn anti-phishing skills to survive. People learn from stories because stories organize events in a meaningful framework and tend to stimulate the cognitive process of the reader [64], [83]. Studies have demonstrated that students in story-based agent conditions perform better in learning than in non-story-based agent conditions [79], [96].

Conceptual–Procedural principle. This principle states that conceptual knowledge and procedural knowledge influence one another in mutually supportive ways and build in an iterative process [60]. In the first version of our game, we taught users specific procedural tips such as “URLs with numbers in the front are generally scams,” or “a company name followed by a hyphen is generally a scam.” We did not teach any conceptual knowledge in the game. Users were able to remember the procedural tips, but without a full conceptual understanding of URLs. Hence, some users applied the lessons learned from the game incorrectly. For example, some users misapplied the rule about IP addresses and thought `www4.usbank.com` was a phishing site because the URL contained the number 4. Other users misapplied the rule “company name followed by hyphen usually means it is a scam” to `web-da.us.citibank.com` (a legitimate site). In the most recent version of our game, we added conceptual knowledge of URLs, explaining the different parts of an URL and which parts are the most important.

We also applied this principle by providing information about how to search for a brand or domain and how to decide which of the search results are legitimate (procedural knowledge) after mentioning that search engines are a good method to identify phishing web sites (conceptual knowledge). In this way, we present conceptual and procedural knowledge iteratively.

5.3.2 Game Description

Here, we describe our game in three parts: story, mechanics, and technology.

Story The main character of the game is Phil, a young fish living in the Interweb Bay. Phil wants to eat worms so he can grow up to be a big fish, but has to be careful of phishers that try to trick him with fake worms (representing phishing attacks). Each worm is associated with a URL, and Phil's job is to eat all the real worms (which have URLs of legitimate web sites) and reject all the bait (which have phishing URLs) before running out of time. The other character is Phil's father, who is an experienced fish in the sea. He occasionally helps Phil out by giving Phil some tips on how to identify bad worms (and hence, phishing web sites).

Mechanics The game is split into four rounds, each of which is two minutes long. In each round, Phil is presented with eight worms, each of which carries a URL that is shown when Phil moves near it (see Figure 1). The player uses a mouse to move Phil around the screen. The player uses designated keys to "eat" the real worms and "reject" the bait. Phil is rewarded with 100 points if he correctly eats a good worm or correctly rejects a bad one. He is slightly penalized for rejecting a good worm (false positive) by losing 10 seconds off the clock for that round. He is severely penalized if he eats a bad worm and is caught by phishers (false negative), losing one of his three lives. We developed this scoring scheme to match the real-world consequences of falling for phishing attacks, in that correctly identifying real and fake web sites is the best outcome, a false positive the second best, and a false negative the worst. The consequences of Phil's actions are summarized in Table 2.

There are four rounds in the game, each one harder than the previous and focusing on a different type of deceptive URL. Table 1 shows the focus of each round. Our implementation selects eight URLs from a pool of twenty for each round, including 12 URLs consistent with the round's focus. The eight URLs illustrate concepts from other rounds to maintain continuity between rounds.

To make the game more engaging and challenging, Phil has to avoid enemy fish while moving around the screen. If Phil comes in contact with an enemy, it eats him and he loses a life. Early

Table 5.1 This table shows the scoring scheme and consequences of the user's actions (through Phil)

	Good worm	**Phishing worm**
Phil Eats	Correct, gains 100 points	False negative, gets phished and loses life
Phil Rejects	False positive, loses 10 seconds	Correct, gains 100 points

Table 5.2 Focus of each round of the game with examples

Round	**Focus**	**Examples**	**Bumper Sticker Message**
1	IP address URLS	http://147.46.236.55/PayPal/	"Don't trust URLs with all numbers in the front"
2	Sub domain URLs	`http://signin.ebay.com.ttps.us`	"Don't be fooled by the word ebay.com in there, this site belongs to ttps.us."
3	Similar and deceptive domains	`http://www.msn-verify.com/` `http://www.ebay-accept.com/`	"A company name followed by a hyphen usually means it is a scam site" "Companies don't use security related keywords in their domains"
4	All previous methods together	eBay sites combining all of above.	

versions of the game included several fast-moving enemies in each round. However, we found that players became distracted by the enemies and had trouble learning. We reduced the number of enemies to one and made them slower so that they did not interfere with learning in later versions of the game.

Players have to correctly recognize at least six out of eight URLs within two minutes to move on to the next round. As long as they still have lives, they can repeat a round until they are able to recognize at least six URLs correctly. If a player loses all three lives the game is over. The game includes brief tutorials before each round and end of round summary, as shown in Figure 3.

Technology The game is implemented in Flash 8. The content for the game, including URLs and training messages, are loaded from a separate data file at the start of the game. This provides us with a great deal of flexibility and makes it easy to quickly update the content. In each round of the game, four good worms and four phishing worms are randomly selected from the twenty URLs in the data file for that round. We also use sound and graphics to engage the user better. This includes sound effects to provide feedback on actions, background music, and underwater background scenes.

5.3.3 Training Messages

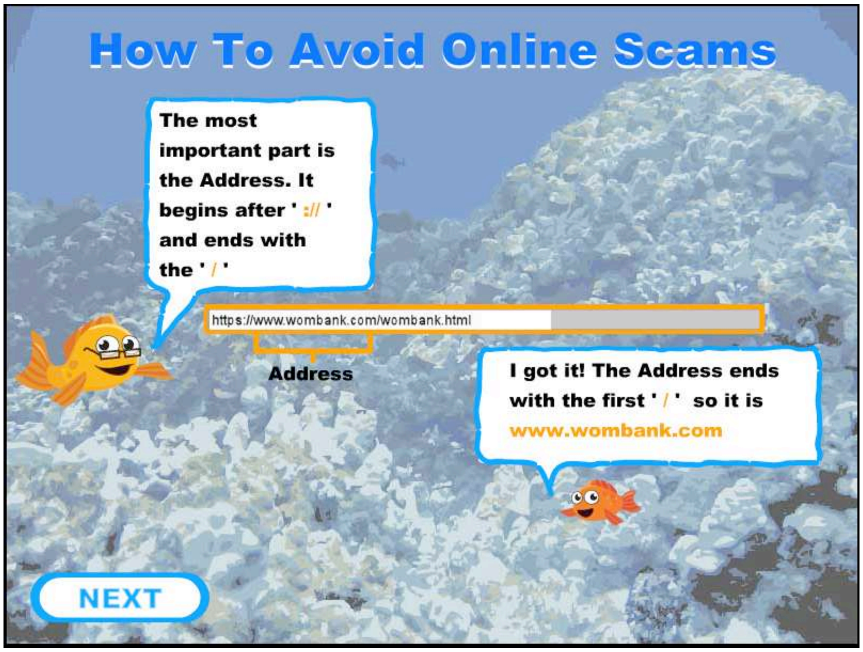
In this section, we discuss details about the training messages that were shown to the users, and the presentation of these training messages.

What to teach Our main focus is to teach users how to identify phishing URLs, where to look for cues in web browsers, and how to use search engines to find legitimate sites.

To teach users to distinguish phishing URLs from legitimate ones, we first sampled a representative list of phishing URLs from the millersmiles.co.uk phishing archive [91], and organized them into three categories: IP-based phishing URLs, long URLs (with sub-domains), and similar and deceptive domains. Next we designed training messages for each type of URL. We iterated on these messages using the philosophy that they should be messages one could place on a bumper

Table 5.3 List of training messages in between rounds; these information helped users to perform better and connect these information with the information presented when they were playing the game.

In Between Round Tip	**# of printed pages**	**Concepts**	**How to do it?**
Tip 1: Don't forget about the URL.	1	- Highlight and point to the address bar in the browser.	
Tip 2: The Middle part of the URL tells you the name of the site.	5	- Highlight the different parts of the URL (Prefix, address and file name).	- Look at the text between the `http://` and the first `/`.The text before the first `/` (this might be with a .com or .org) is the main domain name.
Tip 3: When in doubt, use a search engine!	6	- A search engine is a useful tool to check the legitimacy of a web site.	- Type the domain name or the organization name into Google search engine. The top result is usually legitimate website.
Tip 4: Know the enemies tricks!	1	- Scammers register domains similar to real sites. - They copy logos and contents from real sites to draw you attention. - They request sensitive information. - They point all links to real sites to deceive you.	- Design and logos can be spoofed. Links in the fraudulent website might take to legitimate website.



How To Avoid Online Scams

The most important part is the Address. It begins after ' :// ' and ends with the ' / '

https://www.wombank.com/wombank.html

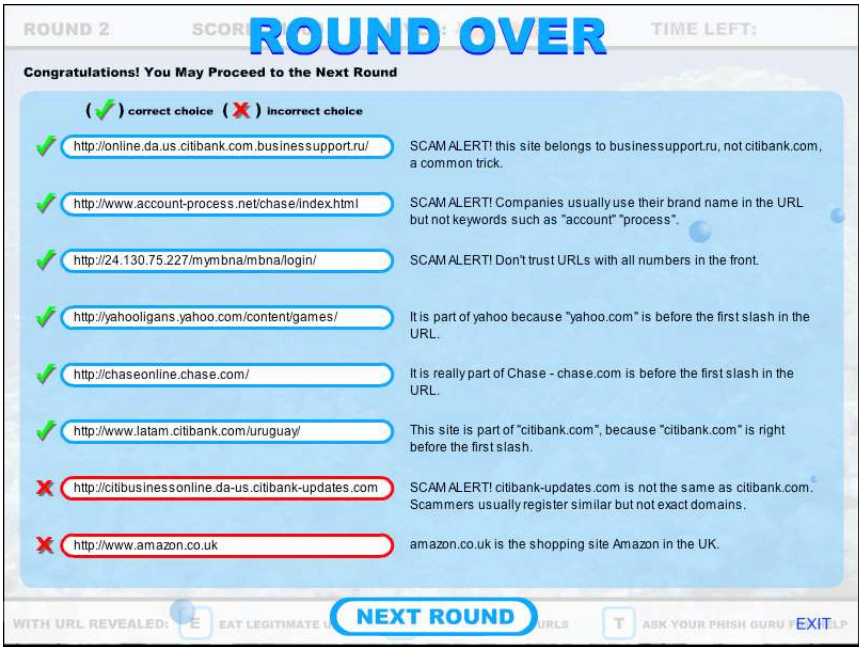
I got it! The Address ends with the first ' / ' so it is www.wombank.com

Figure 5.2 An example training message in between rounds. In this message, Phils father (left) teaches Phil (right) how to identify different parts of a URL and which parts are important.

sticker on a car. For example, for IP-based phishing URLs, we teach “Don’t trust URLs with all numbers in the front.” Table 1 shows a list of bumper sticker messages in the game. To teach users where to look for cues in the browsers, we created a tip that highlighted the browser’s address bar. To teach users how to use search engines to find legitimate sites, we originally used help messages from Phil’s father during the game play. However, as will be discussed in the next section, we found that this was not very effective, so we used a tutorial in between rounds instead.

Where to teach them Training messages are embedded in the following places in the game: (1) feedback during the game, (2) help messages from Phil’s Father during the game, (3) end of the round score sheets, and (4) anti-phishing tips in between rounds.

Feedback during the game: When Phil eats a good URL or rejects a phishing one, we provide some visual feedback such as “yummy” and “got ya” to tell Phil that he got it right. When he eats a phishing URL, he gets phished and is drawn upward by a fishing line and hook. At this point, Phil’s father provides a short tip explaining why the URL is a phishing URL.



ROUND OVER

Congratulations! You May Proceed to the Next Round

(✓) correct choice (✗) incorrect choice

✓	http://online.da.us.citibank.com.businesssupport.ru/	SCAM ALERT! this site belongs to businesssupport.ru, not citibank.com, a common trick.
✓	http://www.account-process.net/chase/index.html	SCAM ALERT! Companies usually use their brand name in the URL but not keywords such as "account" "process".
✓	http://24.130.75.227/mymbna/mbna/login/	SCAM ALERT! Don't trust URLs with all numbers in the front.
✓	http://yahooligans.yahoo.com/content/games/	It is part of yahoo because "yahoo.com" is before the first slash in the URL.
✓	http://chaseonline.chase.com/	It is really part of Chase - chase.com is before the first slash in the URL.
✓	http://www.latam.citibank.com/uruguay/	This site is part of "citibank.com", because "citibank.com" is right before the first slash.
✗	http://citibusinessonline.da-us.citibank-updates.com	SCAM ALERT! citibank-updates.com is not the same as citibank.com. Scammers usually register similar but not exact domains.
✗	http://www.amazon.co.uk	amazon.co.uk is the shopping site Amazon in the UK.

Figure 5.3 "Round over screen". This screen reviews the URLs shown in the round with an indication as to which ones the player identified correctly. The screen also shows a tip to figure out whether the URL is legitimate. This helps provide an opportunity for self-reflection.

Messages from Phil's Father's: Phil can also ask his father for help at any time (by pressing T in the game). His father will provide hints as to what to look for to differentiate good worms from bad ones. Phil's father will also occasionally use a "search engine" and tell Phil the results of the search based on the URL. This is to show Phil how to use a search engine properly to determine the legitimate domain name for a company. This also provides the information players need to determine whether to eat or reject a worm, even if they do not know what the legitimate domain name is for a particular company. In pilot tests of the game, we found that not many users used this option, suggesting that this may not be the most effective way to deliver training.

End of round score sheets: We provide players with an opportunity to reflect on what they learned at the end of each round with a score sheet, as shown in Figure 2. This screen reviews the URLs used in that round, indicates whether or not the player identified each URL correctly, and displays a tip that explains how to figure out whether the URL is legitimate. In our pilot and user study, we found that people often spent a great deal of time on this screen looking over the things they missed. This applies the reflection principle described in Section 3.1.

In previous iterations of the game, we focused solely on teaching people how to discriminate between legitimate and phishing URLs. However, we observed that people needed more scaffolding to help them understand issues like what to look for in the web browser, and how specifically they could use search engines to find real sites. In our current iteration, we added several short tutorials between each round to teach them these kinds of topics. This applies conceptual-procedural principle described in Section 3.1.

5.3.4 Pilot Test

We pilot tested our game with eight users recruited by posting flyers around the Carnegie Mellon University campus. We tested our participants' ability to identify phishing web sites from a set of real and phishing web sites before and after playing the game. The study is a think aloud study where participants talked about strategies they use. The results were encouraging, but highlighted some areas where the game needed improvements.

We found that the game was somewhat effective at teaching users to look at the URL in their browser's address bar when evaluating a web site. Users looked at the address bar when evaluating 14% of the web sites before playing the game and 41% of the web sites after playing the game. The false negative rate decreased from 31% to 17% after users played the game. However, the false positive rate increased from 37% to 48%, in part due to users misinterpreting the URLs they examined.

We observed that users learned some of the URL-related concepts we tried to teach, but not all of them. For example, most users seemed to understand that URLs that have all numbers in the front are usually a sign of scam. However, many users could not properly parse a long URL and did not seem to understand that the most important part of the URL is the right hand side of the domain name. This led them to mis-identify `wellsfargo.com.wfcnet.net` as a legitimate site and `scgi.ebay.com` and `onlineast1.bankofamerica.com` as phishing sites.

We also observed that some users applied the lessons learned from the game incorrectly. For example, some users misapplied the rule about IP addresses (in Table 1) and thought `www4.usbank.com` was a phishing site because the URL contained the number 4. Other users misapplied the rule "company name followed by hyphen usually means it is a scam" to `web-da.us.citibank.com`.

Finally, many participants used wrong strategies to determine the web site legitimacy. For example, one common strategy consisted of checking whether the web site was designed professionally. However, this is not a useful strategy as many phishing sites are exact replicas of professionally designed legitimate sites. Although participants adopted this technique less frequently after the game, some of them still employed a variant of this strategy while using a search engine: they compared the two sites' design (logos, colors) and the exact match of the URL to determine the legitimacy. We believe this is due to users not knowing exactly what to look for to determine web site legitimacy when they use search engines. To summarize, from the pilot test we observed that it is insufficient to teach users how to look for in the URL. We modified our game according to the lessons learned from the pilot testing.

5.3.5 Modified Game

We realized that the initial version of the game focused almost entirely on procedural knowledge. However, some conceptual knowledge about the parts of a URL might have helped users avoid some of the mistakes they made. We added animated messages in between each round of the game to address some of the problems we observed in the pilot study. These messages teach users about the parts of URLs, how to use a search engine to check a suspicious URL, and common tricks used by scam web sites. We designed these messages in a story-like format, in which Phil's father teaches him about URLs at home before he can explore Interweb Bay on his own. Table 3 presents the summary of the training messages that were provided to the user in between rounds, and Figure 4 gives a screenshot of one of the training messages.

5.4 Evaluation 1: Lab Study

In this section, we describe the methodology we used to test the game for its effectiveness in training users in a laboratory study.

5.4.1 Study design

We based the design of our user study on Dhamija et al.'s study, trying to recreate their experiment as much as possible (however, the original materials for Dhamija's study have been lost) [23]. Participants were given the following scenario: "You have received an email message that asks you to click on one of its links. Imagine that you have clicked on the link to see if it is a legitimate web site or a spoofed web site." We then presented participants with ten web sites and asked them to state whether a web site was legitimate or phishing, and to tell us how confident they were in their judgments (on a scale of 1 to 5, where 1 means not confident at all, and 5 means very confident). After evaluating the ten URLs, participants were given fifteen minutes to complete an anti-phishing training task. Finally, participants were shown ten more web sites to evaluate. After finishing this evaluation, participants were asked to complete an exit survey.

We selected twenty web sites (shown in Table 5) to test our participants' ability to identify phishing web sites before and after training. Ten of the sites we selected were phishing sites from popular brands. The other ten were legitimate web sites from popular financial institutions and online merchants, as well as random web sites. We divided the twenty web sites into two groups (A and B), with five phishing sites and five legitimate sites in each group. We randomized the order in which the two groups of URLs were presented so that half the participants saw group A first, and half saw group B first. We hosted the phishing web sites on the local computer by modifying the host DNS file. Thus, our participants were not actually at risk and we were able to show them phishing sites even after they had been taken down.

We told participants that they could use any means to determine a web sites' legitimacy other than calling the company. We also let participants use a separate web browser if they wanted, without prompting them about how or why this might be useful. Some participants used this other web browser to access a search engine to help determine whether a web site was legitimate or not. We used Camtasia Studio to record our participants' computer screens and spoken comments.

We used a between-subjects experimental design to test three training conditions:

1. **Existing training material condition:**In this condition, participants were asked to spend fifteen minutes reading eBay's tutorial on spoofed emails [28], Microsoft's Security tutorial on Phishing [89], the Phishing E-card from the U.S. Federal Trade Commission [32], and a URL tutorial from the MySecureCyberspace portal [97]. We reused the training material condition from our previous study as a control group [70].

2. **Tutorial condition**: In this condition, participants were asked to spend up to fifteen minutes reading an anti-phishing tutorial we created based on the Anti-Phishing Phil game. We include this condition to test the effectiveness of the training messages separate from the game. The tutorial included printouts of all of the between-round training messages. It also included lists of the URLs used in the game with explanations about which were legitimate and which were phishing, similar to the game's end-of-round screens. The 17-page tutorial was printed in color. We designed the tutorial to resemble the game as closely as possible.

3. **Game condition:** In this condition, participants played the Anti-Phishing Phil game for fifteen minutes.

This study was conducted in two phases, separated by five months. For the existing training materials condition, we used data collected during a previous study in September 2006 that measured participants' improvements after reading existing training materials, as compared with a control group that spent fifteen minutes playing solitaire [70]. For the tutorial and game conditions participants were recruited in February 2007 and randomly assigned to these groups. The same procedures were used in September and February for recruiting, screening, and conducting the experiments, although it is possible that the five month delay between the two phases of the experiment introduced some selection bias..

Table 5.4 Participant demographics in each condition

hello	**Existing Training Material**	**Tutorial Group**	**Game Group**
Gender			
Male	29%	36%	50%
Female	71%	64%	50%
Age			
18-34	93%	100%	100%
>34	7%	0%	0%
Education			
High School	14%	7%	7%
College Undergrad	50%	78%	50%
College graduate	14%	7%	21%
Post. Graduate school	21%	7%	21%
Years on the Internet			
3- 5 years	23%	23%	14%
6-10 years	69%	70%	78%
> 11 years	8%	7%	7%

5.4.2 Participant Recruitment and Demographics

In this section, we present the process that we used in recruiting participants for the study; we also describe the demographics of the participants.

We recruited fourteen people for each condition via flyers posted around campus, and with recruitment email on university bulletin boards, and on craigslist.com. We screened participants with respect to their knowledge of computers in general, aiming to recruit only participants who could be considered “non-experts.” We recruited users who answered “no” to two or more of the following screening questions: 1) whether they had ever changed preferences or settings in their web browser, 2) whether they had ever created a web page, and 3) whether they had ever helped someone fix a computer problem. These questions have served as good filters to recruit non-experts in other phishing-related studies [26], [68]. A summary of demographics is shown in Table 4.

5.4.3 Results

In this section, we present the results from the user study. We found that participants in the game condition performed better than the other two conditions in correctly identifying the web sites. We also found that there was no significant difference in false negatives among the three groups. However, the participants in the game group performed better overall than the other two groups.

5.4.3.1 Correlation between Demographics and Susceptibility to Phishing

In this section we present results regarding the correlation between demographics and susceptibility to phishing, user performance, user confidence rating, user feedback, and places where game can be improved.

We found no significant correlation between the participants' performance (measured by total correctnesss) and gender (rho = -0.2, n = 42, p = 0.19), age (spearman rho = 0.008, n = 42, p = 0.96), education (spearman rho = 0.06, n = 42, p = 0.708), race (spearman rho = 0.13, n = 42, p = 0.406), number of hours spent online per week (rho = -0.10, n = 42, p =0.588). Other studies have also found no correlation between these demographics and susceptibility to phishing [23], [140]. The score is positively correlated with years on the Internet (rho = 0.341, n = 42, p = 0.031).

5.4.3.2 User Performance

We measured the effectiveness of user training by examining false positives and false negatives as well as the total percentage of correct sites identified before and after the test. A false positive is when a legitimate site is mistakenly judged as a phishing site. A false negative is when a phishing site is incorrectly judged to be a legitimate site.

Our game condition performed best overall. It performed roughly as well as the existing training material condition in terms of false negatives, and better on false positives. The tutorial condition also performed better than the existing training material in terms of false positives and total correctness. However, these latter results were not statistically significant.

Table 5.5 Percentage of total correct answers for the training group before and after the game

Website	**Real or Fake**	**Description**	**Pre Game % Correct (average confidence)**	**Post Game %correct (average confidence)**
Paypal	Real	Paypal login page	83 (4.6)	100 (4.7)
Bank of America	Real	Bank of America home page; URL: onlin-east.bankofamerica.com	66 (3.5)	100 (4.3)
Wellsfargo bank	Spoof	Faked Wellsfargo home page; layered information request; sub domain deception with URL online.wellsfargo.wfosec.net	83 (3.6)	87 (4.5)
Citibank	Real	Citibank login Page; URL: web-da.us.citibank.com	83 (3.6)	75 (4.5)
Barclays	Spoof	Faked Barclays login page; layered information request; IP address URL	83 (4.2)	100 (4.7)
AOL	Spoof	AOL account update, deceptive domain myaol.com	100 (3.3)	75 (3.4)
Etrade	Real	Etrade home page	100 (4.0)	100 (4.3)
PNC Bank	Spoof	Bank account update; pop-up window over the real PNC Bank web site; security lock; requesting credit card number	66 (4.0)	50 (5.0)
eBay	Real	eBay register page; requesting lots of information	66 (4.2)	62 (4.0)
Halifax Bank	Spoof	Halifax bank login page; deceptive domain halifax-cnline.co.uk.	83 (2.8)	100 (4.5)
Card Financials Online	Real	Card Financial Online (part of MBNA); domain name has nothing to do with MBNA.	50 (3.5)	66 (4.5)
Citicards	Spoof	Citicard account update; lock on the page; requesting a lot of information	50 (4.0)	100 (4.6)
Chase online	Real	Online banking login page; URL: chaseonline.chase.com	100 (4.2)	100 (4.1)
Desjardins	Real	Account login page; unfamiliar foreign bank	50 (3.0)	83 (3.8)
Royal Bank of Canada	Spoof	Sign in online banking page; layered information request; URL has no resemblance with the bank.	37 (4.0)	100 (4.1)
Chase Student	Real	Primitive looking page with few graphics and links	37 (3.0)	66 (3.7)
HSBC	Spoof	Internet banking login page; layered information request; IP address URL	50 (4.0)	100 (5.0)
US Bank	Real	Online banking login page; URL: www4.usbank.com	75 (3.5)	100 (4.6)
eBay	Spoof	Faked eBay login page; IP address URL	75 (3.8)	100 (5.0)
PayPal	Spoof	Fake URL bar displaying the real Paypal URL; not requesting much information	50 (3.2)	0 (4.0)

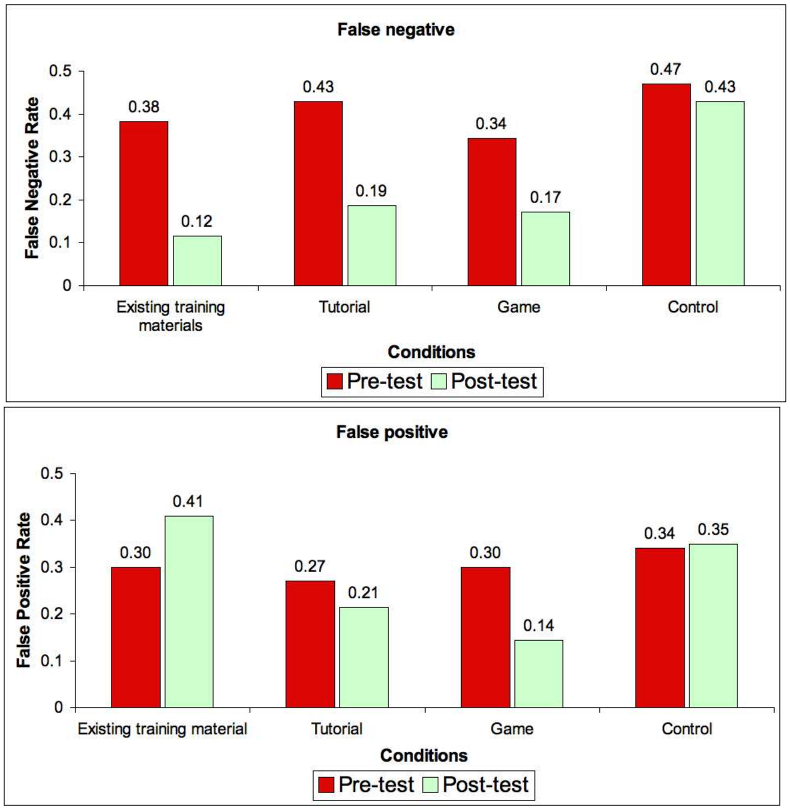


Figure 5.4 User performance in the experimental conditions: existing training materials, tutorial only, game, and control condition. N=14 in all conditions. The graph on the left shows false negative rates.The existing training material performed best on false negatives. However, the difference is not statistically significant. The graph on the right shows False Positive Rate. The false positives increased in the existing materials condition, and decreased in both the tutorial and game condition, with the game condition showing the highest reduction.

Post test false negative rates in all three groups decreased significantly from the pre test values. For the existing training materials condition, the false negative rate fell from 0.38 to 0.12 (paired t-test: 1=0.38, 2=0.12, $p = 0.01$); for the tutorial condition, it changed from 0.43 to 0.19 (paired t-test: 1=0.43, 2=0.19, $p < 0.03$); for the game condition, it changed from 0.34 to 0.17 (paired t-test: 1=0.34, 2=0.17, $p <0.02$). There is no statistical difference between the three groups in either the pre test (oneway ANOVA, F(2,41)=0.52, p=0.60), or post test (oneway ANOVA, F(2,41)=0.81, p=0.45). These results are shown in Figure 5.

Post test false positive rates decreased significantly in the game condition (paired t-test: 1=0.30, 2=0.14, $p < 0.03$). The one-way ANOVA revealed that false positive rates differed significantly in the post test ($F(2, 41) = 4.64$, $p < .02$). The Tukey post-hoc test revealed that the game condition has significantly lower false positives than the existing training materials. No other specific post-hoc contrasts were significant. The results are shown in Figure 6.

Combining false positive and false negatives we derived a measure for the total correctness. We found in the post test that the game condition performed better than the existing training material condition (2 sample t test, $p<0.02$). We did not find the tutorial condition improvement to be significant over the existing training material condition; however, this is likely due to our small sample size. These results are shown in Figure 7.

False negative rates. N = 14 in all conditions. The existing training material performed best on false negatives. However, the difference is not statistically significant.

False Positive Rate. N = 14 in all conditions. The false positives increased in the existing materials condition, and decreased in both the tutorial and game condition, with the game condition showing the highest reduction.

Total correctness for the test groups. N = 14 in all conditions. The game condition shows the greatest improvements.

5.4.3.3 User Confidence Rating

Users became more confident about their judgments after the game or the tutorial conditions. We did not observe the existing training material improving user confidence in a statistically significant way.

The average user confidence rating in the game condition increased from 3.72 (variance = 0.09) to 4.42 (variance = 0.10). This change is statistical significant (paired t –test, $p < 0.001$). In contrast, user confidence in the existing training material condition did not improve in a statistically significant way: the average confidence rating was 4.18 pre test (variance = 0.18) and 4.32 post test (variance = 0.15).

5.4.3.4 User Feedback

In the post test, we asked participants to measure on a 5- point Likert scale how much they felt they had learnt and how important was the information they learnt. Ninety-three percent of the users either agreed or strongly agreed that they had learned a lot (u = 4.21, std = 0.58), and 100% of them agreed or strongly agreed that they had learned a lot of important information (u = 4.36 std=0.50). On a five point scale, we also asked them to rate the educational and fun levels of the game. Ninety-three percent of the user felt the educational value of the game was very good or excellent (u=4.28, var = 0.61). Fifty percent of the users considered the fun level of the game as very good or excellent (u = 3.7 var = 0.78).

We asked similar questions about educational value and fun level in the existing training material condition. Ninety-three percent of the users also felt the educational value of the existing training material was very good or excellent (u=4.28 var = 0.59), where as only twenty-nine percent of the users considered the fun level of the existing training materials to be very good or excellent (u = 2.8 var = 1.36).

5.4.3.5 Where the Game is Failing

We found that users in the game group and the tutorial group performed worse when examining two websites. The first website is a fake address bar attack, where we showed users a Paypal website with the address bar spoofed. Six of the users in the game condition were unable to identify this attack in the post test, whereas only three users in the existing training material condition fell for it. We hypothesize that users are more prone to this kind of attacks because, after the training, they look specifically for clues in the URL, and if the clues confirm their belief, they do not look further. (Luckily, current browsers now address this kind of vulnerability.)

Two users also fell for the “similar domain attack” after the game condition, in which we showed them myaol.com for account updates. This is an easy attack to identify if users notice the large amount of information requested, because of this reason, none of the users fall for it in the pre test. This problem highlights two lessons: first, some users still have problems with phishing domains that are similar to the real ones; second, they tend to look less for other clues other than the URL, and if the URL does not raise suspicion, they do not look further.

5.4.3.6 Effect of Training

Security education plays an important role in increasing users' alertness towards security threats. Alert users are cautious, and less likely to make mistakes that will leave them vulnerable to attack (false negatives). However, cautious users tend to misjudge non-threats as threats (false positives) unless they have learned how to distinguish between the two. Thus, good user security education should not only increase users' alertness, but also teach them how to distinguish threats from non-threats. In this section we use signal detection theory (SDT) [78, 120] to quantify the ability to discern between signal (phishing websites) and non-signal or noise (legitimate websites).

We use two measures: *sensitivity* (d') and *criterion* (C). In our user studies, we define sensitivity to be the ability to distinguish phishing websites from legitimate websites, which is measured by the distance between the mean of signal and non-signal distributions. The larger the value of d', the better the user is at separating signal from noise. *Criterion* is defined as the tendency of users towards caution when making a decision. More cautious users are more likely to have few false negatives and many false positives, while less cautious users are likely to have many false negatives and few false positives. Figure 5.5 shows example distributions of user decisions about legitimate and phishing websites. The criterion line divides the graph into four sections representing true positives, true negatives, false positives, and false negatives. Training may cause users to become more cautions, increasing C and moving the criterion line to the right. Alternatively, training may cause users to become more sensitive, separating the two means. In some cases training may result in both increased caution and increased sensitivity or in decreased caution but increased sensitivity.

We calculated C and d' for the participants in our user study, Table 5.6 presents the results. We found that in the existing training material condition, the sensitivity increases from 0.81 in pre test to 1.43 in post test. This increase is significant ($p < 0.05$). We also found that users became cautious after the training, as the d' changes from 0.03 in pre test to -0.51 in post test ($p < 0.025$). This result (users becoming more cautious) was also shown by Jackson et. al [2]. In contrast

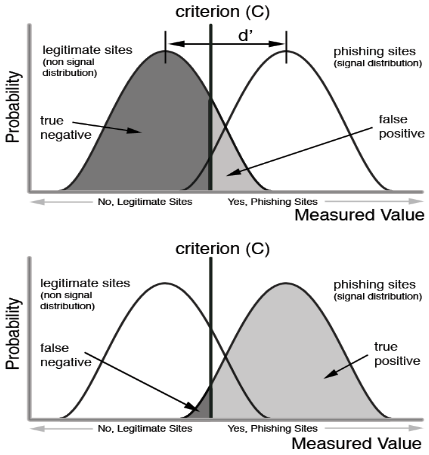


Figure 5.5 Applying signal detection theory (SDT) to anti-phishing education We treat legitimate websites as “non signal,” and phishing websites as “signal.” Sensitivity (d’) measures users’ ability to distinguish signal from non-signal. Criterion (C) measures users’ decision tendency (C < 0 indicates cautious users , C = 0 indicates neutral users, C > 0 indicates liberal users). As a result of training users may a) become more cautious, increasing C; b) become more sensitive, increasing d’; or c) a combination of both.

to the existing training material condition, the sensitivity increased from 0.93 to d’_post = 2.02 (p<0.025) in the game condition. Also, the decision criterion did not change significantly (C_pre = 0.06, C_post = 0.06) in the game condition. This shows that the improvement in the performance is due to learned ability to better distinguish phishing websites and real websites.

Table 5.6 Results from the Signal Detection Theory analysis. This shows that users had a greater sensitivity with Anti-Phishing Phil, meaning that they were better able to distinguish between phishing and legitimate sites. Consequently, users were able to make better decisions in the game condition compared to the users becoming conservative in the other condition.

	Sensitivity (d’)			Criterion (C)		
	Pre test	Post test	Delta	Pre test	Post test	Delta
Existing training materials	0.81	1.43	0.62 *	0.03	-0.51	-0.54 **
Anti-phishing Phil	0.93	2.02	1.09 **	0.06	0.06	0
* p <0.05, ** p < 0.025						

5.5 Evaluation 2: Anti-Phishing Phil Field Study

In this section, we discuss new results from data we collected in a real-world deployment of Anti-Phishing Phil. Our results provide more evidence that Anti-Phishing Phil is effective for knowledge acquistion and knowledge retention.

5.5.1 Study design

We recruited participants for an online study through online mailing lists postings offering participants a chance to win a raffle for a $100 Amazon gift certificate. We used a between-subjects design to test two conditions. In the control condition, participants saw 12 websites and were asked to identify whether each website seen was phishing or not. After doing this, the participants were taken to the game. In the game condition, participants were shown six websites before playing the game (pre-test) and another six websites after they finished playing the game (immediate post-test). To measure retention, we emailed participants seven days later and asked them to take a similar test (delayed post-test). In total, we tested each participant in the game condition on 18 websites divided into three groups of three phishing websites and three legitimate websites. We randomized the order of websites within each group, and the order in which the groups were shown to each participant.

5.5.2 Participants

Over the course of two weeks (Sep 25, 2007 to Oct 10, 2007), 4,517 people participated in the study. In the game condition, 2,021 users completed both pre-test and immediate post-test, 674 of whom also came back one week later for the delayed post-test. In our analysis we focus on people who completed pre-test, immediate post-test, and delayed post-test. We had 2,496 participants in the control condition. Among the total participants, there were 78% male, 15.6% female, and 6.4% did not give their gender; 4.8% were 13 - 17 years old, 43.7% were 18 - 34 years old, 44.3% were 35 - 64 years old, 0.5% were more than 65 years, and 6.8% did not provide their age.

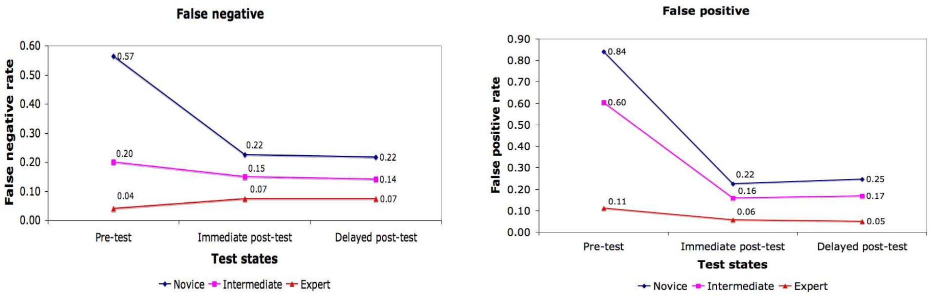


Figure 5.6 False negative and false positive for Anti-Phishing Phil in the real-world. Novice users show greatest improvement in false negative and false positive

5.5.3 Results

Our results demonstrate that users are able to more accurately and quickly distinguish phishing websites from legitimate websites after playing the game, and that they retain knowledge learned from the game for at least one week.

We classified the game condition participants into three categories based on their pre-test scores: novice (0 - 2 correct), intermediate (3 - 4 correct) and expert (5 - 6 correct). As illustrated in Figure 5.6, novice users showed the greatest improvement, with false positive rate decreasing from 42% to 11.2% (paired t-test, $p < 0.0001$), and false negative rate decreasing from 28.3% to 11.2% (paired t-test, $p < 0.0001$). The intermediate group also showed statistically significant improvements, although not as large as the novice group. Finally, we did not observe any statistically significant improvements for the expert group. Delayed post-test scores did not decrease from immediate post-test scores; demonstrating that participants retained their knowledge after one week.

Participants were able to determine website legitimacy more quickly after playing the game. The mean time users in the game group took to determine a webwebsite's legitimacy before the game was 21.2 seconds. After the game, it decreased to 11.2 seconds (paired t-test, $p < 0.0001$). The mean scores for the control group does not change in a statistically significant way (pre - 18.5 seconds, post - 18.6 seconds).

Those who did not come back for the delayed post-test performed slightly worse than those who did come back. Their immediate post-test score is 83.8% for those who did not come back and 89.1% for those who did come back one week later (two sample t-test, $p < 0.001$). One possible explanation is that those who were more confident in their performance were more likely to come back. To validate this hypothesis, we conducted a Chi-square test of the percentage of novice, intermediate and expert users completed the immediate post-test, or delayed post-test. We found that there were more experts and fewer intermediate and novices in the delayed post-test group ($p < 0.001$).

Before playing the game mean accuracy scores for males were significantly higher than for females (males = 75.5%, females = 64.4%, two sample t-test, t = 8.48, $p < 0.0001$). However, the two groups improved similarly after playing the game (two proportion test, 14.2% versus 12.4%, p = 0.192). There was also a significant difference in pre-test performance between different age groups (one way ANOVA $F = 7.29$, $p < 0.01$). A Turkey simultaneous 95% confidence interval test reveals that participants whose age is less than 18 performed worse than those who are between 18 and 64. There is no statistical difference in performance between the ages groups 18-35 and 36-64. We observed similar trends in immediate post-test performance (one way ANOVA, $F = 23.05$, $p < 0.01$). These results suggest that teenagers may be particularly susceptible to phishing attacks. The mean scores for the age group 13-17 years was 3.9 while the mean score was 4.6 for both 18-34 and 35-64 age groups.

We used the data from the game to determine which types of URLs are most difficult for people to identify correctly. Especially challenging ones are the URLs longer than the address bar and deceptive URLs that look similar to legitimate URLs with some added text (e.g. http://www.msn-verify.com/). The more challenging the URL, the more likely game players are to use the game's help feature ($r = -0.645$, $p < 0.001$). From the game data, we found that users are most confused with long URLs. This confusions makes them susceptible to sub-domain attacks such as (https://citibusinessonline.da-us.citibahnk.com/cbusol/signon.do). Users are also confused with very similar URLs. For example, www.citicards.net (as opposed to www.citicards.com), www.eztrade.com (as opposed to www.etrade.com). This suggests for further investigation on ways to teach to remove these confusions among users.

5.5.3.1 Effect of Training

Using the Signal detection method that we introduced in section 5.4.3.6. We calculated C and d' for our evaluation of existing online training materials, PhishGuru retention and transfer study, Anti-Phishing Phil laboratory study, and Anti-Phishing Phil field study, as summarized in Table 5.7. We found that after reading existing training materials, users became significantly more cautious without becoming significantly more sensitive. Thus these materials serve to increase alertness, but do not teach users how to distinguish legitimate websites from fraudulent ones. After playing Anti-Phishing Phil, users became both significantly more sensitive and liberal, indicating that performance improvements from playing the game are due to learning. (Note, in the laboratory study we did not observe the Criterion change that we observed in the field study.) PhishGuru embedded training increased both sensitivity and caution, but these results are not statistically significant due to the small number of user decisions considered in the analysis. The pre-test Criterion for the existing training and Anti-Phishing Phil studies indicate these users started off more cautious than those in the PhishGuru study. This is likely due to the fact that users were primed to think about security in the former studies and not in the latter study.

Table 5.7 Signal Detection Theory analysis. PhishGuru and Anti-Phishing Phil increased user's sensitivity, while existing training materials made users more cautious. * indicates statistically significant differences ($p < 0.05$).

	Sensitivity (d')			**Criterion (C)**		
	Pre-test	**post-test**	**Delay**	**Pre-test**	**post-test**	**Delay**
Existing training materials	0.81	1.43	–	0.03	-0.51*	–
Anti-Phishing Phil laboratory study	0.93	2.02*	–	0.06	0.06	–
Anti-Phishing Phil field study	1.49	2.46*	2.47	-0.35	0.02*	0.0

Conclusions and Future Work

In this paper, we presented the design and evaluation of Anti-Phishing Phil, a game that teaches users not to fall for phishing attacks. Our objective in developing the anti-phishing game was to teach users three things: (1) how to identify phishing URLs, (2) where to look for cues in web browsers, and (3) how to use search engines to find legitimate sites. In particular, the game teaches users about identifying three types of phishing URL's: IP based URLs, sub domain, and deceptive.

We conducted two user studies. In the first study, we compared the effectiveness of the game with existing online training materials and a tutorial we created based on the game. We found that participants who played the game performed better at identifying phishing websites than participants who completed the two other types of training. Using signal detection theory, we also showed that while existing online training materials increase awareness about phishing (which can help people avoid attacks), our game also makes users more knowledgeable about techniques they can use to identify phishing web sites.

In the second study, we tested Phil from data we collected in a real-world deployment of Anti-Phishing Phil. Our results provide more evidence that Anti-Phishing Phil is effective for knowledge acquistion and knowledge retention.

Our results show that interactive games can be a promising way of teaching people about strategies to avoid falling for phishing attacks. Our results suggest that applying learning science principles to training materials can stimulate effective learning. Finally, our results strongly suggest that educating users about security can be a reality rather than just a myth [48].