Teaching Johnny Not to Fall for Phish

PONNURANGAM KUMARAGURU, STEVE SHENG, ALESSANDRO ACQUISTI,
LORRIE FAITH CRANOR, and JASON HONG
Carnegie Mellon University

Phishing attacks, in which criminals lure Internet users to websites that spoof legitimate websites, are occurring with increasing frequency and are causing considerable harm to victims. While a great deal of effort has been devoted to solving the phishing problem by prevention and detection of phishing emails and phishing websites, little research has been done in the area of training users to recognize those attacks. Our research focuses on educating users about phishing and helping them make better trust decisions. We identified a number of challenges for end-user security education in general and anti-phishing education in particular: users are not motivated to learn about security; for most users, security is a secondary task; it is difficult to teach people to identify security threats without also increasing their tendency to misjudge non-threats as threats. Keeping these challenges in mind, we developed an email-based anti-phishing education system called "PhishGuru" and an online game called "Anti-Phishing Phil" that teaches users how to use cues in URLs to avoid falling for phishing attacks. We applied learning science instructional principles in the design of PhishGuru and Anti-Phishing Phil. In this paper we present the results of PhishGuru and Anti-Phishing Phil user studies that demonstrate the effectiveness of these tools. Our results suggest that, while automated detection systems should be used as the first line of defense against phishing attacks, user education offers a complementary approach to help people better recognize fraudulent emails and websites.

Categories and Subject Descriptors: D.4.6 [**Operating Systems**]: Security and Protection; H.1.2 [**Models and Principles**]: User/Machine Systems; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms: Design, Experimentation, Security, Human factors

Additional Key Words and Phrases: Embedded training, learning science, instructional principles, phishing, email, usable privacy and security, situated learning

1. INTRODUCTION

A semantic attack is a computer-based attack that exploits human vulnerabilities. Rather than taking advantage of system vulnerabilities, semantic attacks take ad-

This work was supported by the National Science Foundation grant number CCF-0524189; by the Army Research Office grant number DAAD19-02-1-0389; and by the Fundao para a Cincia e Tecnologia (FCT), Portugal, under a grant from the Information and Communications Technology Institute (ICTI) at CMU.

Author's addresses: P. Kumarauguru, Carnegie Mellon University; email: ponguru@cs.cmu.edu; S. Sheng, Carnegie Mellon University; email: shengx@cmu.edu; A. Acquisti, Carnegie Mellon University; email: acquisti@andrew.cmu.edu; L. Cranor, Carnegie Mellon University; email: lorrie@cs.cmu.edu; J. Hong, Carnegie Mellon University; email: jasonh@cs.cmu.edu.

vantage of the way humans interact with computers or interpret messages [Schneier 2000]. Since 2003, we have seen a dramatic increase in a semantic attack known as *phishing*, in which victims get conned by spoofed emails and fraudulent websites. Phishing attacks take advantage of users' inability to distinguish legitimate company websites from fake ones. Phishers send out spoofed emails that look as if they had been sent by trusted companies. These emails lead to spoofed websites that are similar or virtually identical to legitimate websites, and lure people into disclosing sensitive information. Phishers use this information for criminal purposes such as identity theft, financial fraud, and corporate espionage [James 2005; Lininger and Vines 2005].

Developing countermeasures for phishing is a challenging problem because victims are helping attackers by giving away their credentials. It is also difficult to detect phishing websites and emails because they often look legitimate. Finally, users frequently ignore phishing warning messages from anti-phishing tools [Florencio and Herley 2005; Egelman et al. 2007; Wu et al. 2006].

A variety of strategies to protect people from phishing have been proposed. These strategies fall into three major categories: *silently eliminating the threat*, by finding and taking down phishing websites, as well as by detecting and deleting phishing emails automatically; *warning users about the threat*, through toolbars, browser extensions, and other mechanisms; and *training users not to fall for attacks*. We argue that these approaches are complementary. Specifically, automated detection systems should be used as the first line of defense against phishing attacks, but since these systems are unlikely to perform flawlessly, they should be complemented with user interfaces and user education to help people better recognize fraudulent emails and websites.

Most anti-phishing research has focused on solving the problem by eliminating the threat or warning users. However, little work has been done on educating people about phishing and other semantic attacks. Educating users about security is challenging, particularly in the context of phishing, because: (1) users are not motivated to read about security in general and therefore do not take time to educate themselves about phishing; (2) for most users, security is a secondary task (e.g. one does not go to an online banking website to check the SSL implementation of the website, but rather to perform a banking transaction); and (3) it is difficult to teach people to make the right online trust decision without increasing their tendency to mis-judge non-threats as threats.

In this paper, we present and extend our cumulative research on anti-phishing education. We examine the questions of what to teach people, how to present the training, and how to motivate users to pay attention to the training. In Section 2, we present background and related work on combatting phishing. In Section 3, we present an overview of learning science, highlighting principles designed to help people acquire and retain knowledge better. In Section 4, we present our analysis of existing web-based anti-phishing training materials, to determine their strengths and weaknesses and to identify important training messages. The results from this analysis helped inform the design of two tools we developed to educate people about phishing. We continue in Section 5 with an overview of the design and evaluation of PhishGuru, the first of these tools. PhishGuru is an embedded training system

that we developed to teach people about phishing during their normal use of email. Using lab experiments, we showed that embedded training works better than the current practice of emailing security notices [Kumaraguru et al. 07a]. We also demonstrated that users retain what they learn from PhishGuru for at least one week [Kumaraguru et al. 07b]. In Section 6, we present the design and evaluation of the second of these tools, Anti-Phishing Phil, an online game that teaches people how to avoid phishing attacks. We found that participants who played the game were better at identifying fraudulent websites compared to participants who did not play the game [Sheng et al. 2007]. We also present the results of a new study in which over 4,500 people played Anti-Phishing Phil. This study demonstrated that people are able to distinguish phishing websites more accurately and quickly after playing our game and could retain knowledge learned from the game for at least one week. In Section 7, we discuss the effect of training by applying Signal Detection Theory (SDT) to our results. Finally, in Section 8, we discuss our conclusions about anti-phishing user education.

2. BACKGROUND AND RELATED WORK

The various strategies to protect people from phishing fall into three major categories: silently eliminating the threat, warning users about the threat, and training users not to fall for attacks. These categories of anti-phishing strategies mirror the three high-level approaches to usable security discussed in the literature: build systems that "just work" without requiring intervention on the part of users, make security intuitive and easy to use, and teach people how to perform security-critical functions [Cranor 2008].

2.1 Silently Eliminating the Threat

The premise of silently eliminating the phishing threat is to protect users without requiring any awareness or action on their part. When phishing web sites are blocked or taken down, phishing emails are deleted before they reach the recipient's inbox, and the perpetrators of phishing attacks are arrested, users are protected.

A variety of efforts are aimed at identifying phishing email and websites so that they can be blocked and phishing URLs can be added to blacklists. Some email providers use spam filters to identify phishing emails and verify the domain of the sender [Sender Policy Framework 2006; Yahoo 2007]. Both blacklist and machine learning techniques can also be used to detect phishing emails [Chandrasekaran et al. 2006; Fette et al. 2006; Abu-Nimeh et al. 2007].

If phishing could be completely eliminated using these methods, there would be no need for other protection strategies. However, existing tools are unable to detect phishing emails and phishing websites with one hundred percent accuracy. For example, in a 2007 study, even the best anti-phishing toolbars missed over 20% of phishing websites [Zhang et al. 2007] and a 2009 study found that most anti-phishing tools did not start blocking phishing sites until several hours after phishing emails had been sent luring users to those sites [Sheng et al. 2009].

2.2 Warning Users About the Threat

There are also a number of tools that warn users that the website they are visiting is likely to be fraudulent, either by providing explicit warnings or by providing

interfaces that help people notice that they may be on a phishing website. For example, researchers have proposed user interfaces for "trusted paths" that assist users in verifying that their browser has made a secure connection to a trusted website [Dhamija and Tygar 2005; Ye and Smith 2002]. In addition, the major web browsers now warn users who attempt to visit phishing websites and several web browser toolbars provide cues as to the legitimacy of a website [Account Guard 2006; Netcraft 2006; SpoofGuard 2006; SpoofStick 2006].

However, these approaches have significant weaknesses. In particular, these tools require user involvement by end-users, and thus are unlikely to be effective if they are not extremely simple to understand and use. User studies have shown that users often do not understand or act on the cues provided by toolbars [Miller and Wu 2005; Wu et al. 2006]. Passive indicators that do not interrupt the task are especially problematic; active indicators that block a phishing web page have been shown to be significantly more effective [Egelman et al. 2007].

2.3 Training Users not to Fall for Attacks

The core idea in a third thread of work—and the main thrust of the work presented in this paper—is that users can be trained to actively protect themselves from phishing threats. It is worth noting, however, that some experts argue that "security user education is a myth" [Gorling 2006]. Others have commented that education is not a feasible solution for phishing and other security attacks because security education "puts the burden on the wrong shoulder" [Nielsen 2004] and security is a secondary goal for users [Evers 2006]. Furthermore, evaluations of some security-related educational materials have found these materials to be ineffective [Anandpara et al. 2007; Jackson et al. 2007]. In general, we found that existing online anti-phishing training materials tend to make users more cautious about opening and acting upon email, but do not teach people how to determine whether a website or email is fradulent (See Section 4). Hence, there is a need for developing training materials specifically about identifying phishing and semantic attacks to teach users to make better online trust decisions.

Some experts argue that education and training can prevent users from falling for phishing and other attacks [Emigh 2005], [Cranor and Garfinkel 2005, Chapter 14], [Jakobsson and Myers 2006, Chapter 3], [Hight 2005] and research has shown that education can be an effective solution to the phishing problem [Ferguson 2005]. Many companies spend a considerable amount of money in educating their employees about security [Gordon et al. 2006]. Security education is useful not only for teaching security skills, but also for motivating the need for security. Studies have also shown that the majority of computer users are security conscious, as long as they perceive the need for secure behaviors [Adams and Sasse 1999, pp. 45]. Phishing education can also be conducted in a classroom setting with good results [Robila and Ragucci 2006]. However, it is difficult to train large number of users through classroom sessions alone.

A variety of anti-phishing educational materials are also available online, for example, from the Federal Trade Commission [Federal Trade Commission 2006a; 2006b], the Anti-Phishing Working Group [Anti-Phishing Working Group 2007], and from companies targeted by phishers [eBay 2006; Microsoft Corporation 2006]. However, users seldom seek out these materials and, as we have observed in our

studies, users tend to ignore emails directing them to these materials. As others have noted, most users are unlikely to spend much time reading security-related tutorials, and "gentler methods for providing users with the right guidance at the right time" are necessary [Whitten 2004].

A more interactive approach is to provide web-based tests that let users assess their own knowledge of phishing. For example, Mail Frontier has set up a website containing screenshots of potential phishing emails [Mail Frontier 2006]. Users are scored based on how well they can identify which emails are legitimate and which are not. However, while this approach raises awareness about phishing, a user study found that it is not an effective training method [Anandpara et al. 2007]. Nonetheless, the idea of integrating self tests with other anti-phishing training materials warrants further examination.

Another method for educating users is to send fake phishing emails to test users' vulnerability, and then follow up with training. Subsequent fake phishing emails can be used to measure improvements in phishing detection abilities. This approach has been used with students [Jagatic et al. 2007; Ferguson 2005], as well as with employees [New York State Office of Cyber Security & Critical Infrastructure Coordination 2005] and has shown that education can improve participants' ability to identify phishing emails. Researchers have also looked at non-traditional approaches to security education, such as comic strips [Jakobsson 2007].

Our work differs from past work in that we are focused on understanding what educational approaches are effective in teaching people about phishing and actually protecting them in practice. We explore which concepts to teach and how to motivate users to read and understand training materials. Furthermore, our work measures knowledge retention and knowledge transfer to evaluate the effectiveness of training.

3. LEARNING SCIENCE

In this section, we provide an overview of the instructional design principles and methods we drew from the field of learning science, the body of research that examines how people gain knowledge and learn new skills. These principles and methods informed the design and evaluation of our training mechanisms, discussed later in this paper.

3.1 Instructional design principles

Education researchers have developed instructional design principles to guide the development of effective educational materials. Table I summarizes the instructional design principles that we used in our training materials. We selected these principles as they are the most powerful of the basic instructional design principles applicable to online security training.

—*Learning-by-doing principle*: ACT-R (Adaptive Control of Thought–Rational) was developed to model human cognition and learning. One of the fundamental hypotheses of ACT-R is that knowledge and skills are acquired and strengthened through actual practice [Anderson 1993]. Experiments with cognitive tutors have shown that students who practice skills that they have just learned perform better than students who do not [Aleven and Koedinger 2002; Eberts 1997; Schmidt and

Table I. Instructional design principles used in our research

Principle	Explanation
Learning-by-doing	People learn better when they practice the skills they are learning
Immediate feedback	Providing immediate feedback during the knowledge acquisition phase results in efficient learning
Conceptual-procedural	Conceptual and procedural knowledge influence one another in mutually supportive ways and build in an iterative process
Contiguity	Presenting words and pictures contiguously (rather than isolated from one another) enhances learning
Personalization	Using conversational style rather than formal style enhances learning
Story-based agent environment	Using characters in a story enhances learning
Reflection	Presenting opportunities for learners to reflect on the new knowledge they have learned enhances learning

Bjork 1992].

—*Immediate feedback principle*: Researchers have shown that providing immediate feedback during the knowledge acquisition phase results in efficient learning, guidance towards correct behavior, and a reduction in unproductive floundering [Anderson et al. 1995; Mathan and Koedinger 2003; Schmidt and Bjork 1992]. Corbett et al. showed that students who got immediate feedback performed significantly better than students who got delayed feedback [Corbett and Anderson 2001].

—*Conceptual-procedural principle*: A *concept* is a mental representation or prototype of objects or ideas (e.g. phishing) [Clark 1989]. A *procedure* is a series of clearly defined steps which results in the achievement of a given task (e.g. logging onto a computer) [Clark 1989]. The conceptual-procedural principle states that conceptual and procedural knowledge influence one another in mutually supportive ways and build in an iterative process [Johnson and Koedinger 2002]. For example, in an experiment about teaching decimal places, students who were presented with concepts and then procedures in an interleaved fashion performed better than students who were first presented with all of the concepts and then presented with all of the procedures [Koedinger 2002; Johnson and Koedinger 2002].

—*Contiguity principle*: Mayer et al. proposed 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 [Mayer and Anderson 1992]. In an experiment, students who learned about lightning storms performed better when words and pictures were close to each other (*spatial-contiguity*) [Moreno and Mayer 1999].

—*Personalization principle*: This principle states that: using conversational style rather than formal style enhances learning [Clark and Mayer 2002]. People are more likely to try to understand the instructional material if it is presented in a way that makes them feel that they are part of a conversation rather than just receiving information. Researchers suggest using "I," "we," "me," "my," "you," and "your" in instructional materials to enhance learning [Mayer 2001]. In an experiment aimed at teaching arithmetic order-of-operation rules, students who

received conversational style messages were more engaged and learned more than the control group [Cordova and Lepper 1996].

—*Story-based agent environment principle*: Agents are characters who help guide learners. These characters can be represented visually or verbally and can be cartoon-like or real life characters. People are more motivated to learn when guided by an agent. Learning is further enhanced if the materials are presented within the context of a story [Mayer 2001].

—*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 systems include opportunities for learners to reflect on the new knowledge they have learned [Committee on Developments in the Science of Learning and National Research Council 2000].

3.2 Measuring learning

Effective education should help learners acquire new knowledge (*knowledge acquisition*), perform the skills learned in the long run (*knowledge retention*), and apply learned knowledge to related tasks (*knowledge transfer*) [Schmidt and Bjork 1992]. In this paper, we operationalize these three requirements as follows:

—*Knowledge Acquisition* (KA): the ability to process and extract knowledge from instructional materials. Learners should be able to use the acquired knowledge to make decisions [Mandl and Levin 1989; Bahrick 1979]. This is usually evaluated by asking people to repeat or apply knowledge just after learning.

—*Knowledge Retention* (KR): the ability to retain or recall knowledge after some time has passed from the original time of knowledge acquisition. Researchers have proposed a variety of approaches to quantifying retention [Rubin and Wenzel 1996].

—*Knowledge Transfer* (KT): the ability to apply the knowledge gained in one situation to a different situation after some time has passed from the time of knowledge acquisition. The definition and measurement of transfer are heavily debated in the learning science literature [Bransford and Schwartz 2001; Singley and Anderson 1989; Schwartz and Bransford 1998]. Researchers have developed a taxonomy to classify different types of transfers [Gagne et al. 1948; Barnett and Ceci 2002]. Two types of transfers that are frequently discussed are immediate (near) transfer and delayed (far) transfer [Fong and Nisbett 1991; Merrienboer et al. 1997]. Near transfer is transfer of knowledge and skills from one context to a closely related context, while far transfer is transfer of skill from one context to an entirely different context.

4. EVALUATION OF EXISTING ONLINE TRAINING MATERIALS

In Section 2, we discussed user studies that measured the effectiveness of specific training materials. However, those studies did not analyze the quality of the training materials being tested and did not consider ways of designing more effective training materials. In this section, we present our analysis of online training materials using the instructional design principles described in section 3.1, as well as the results of a user study examining the effectiveness of existing training materials.

4.1 Training material selection and analysis

We compiled a list of 25 online anti-phishing training materials and consulted with experts to select a set of frequently mentioned guidelines that offered simple and effective steps that end users could take to avoid falling for phishing attacks. We eliminated guidelines that focussed on strategies that would be difficult for many users, such as using networking tools to determine the age and owner of a domain. Based on this analysis, we selected the following five guidelines: (1) Never click on links in emails; (2) Always access a website by typing in the real website address into the web browser; (3) Never trust phone numbers in emails — look up phone numbers using a reliable source such as a phone directory or credit card statement; (4) Never respond to emailed requests for personal information; and (5) Be suspicious of website that ask for too much personal information.

The first guideline was somewhat controversial among the experts we consulted. While they agreed that users who do not click on links will not be susceptible to most email-based phishing attacks, some experts argued that email links offer considerable convenience and value to users. As such, they argued that it would be unrealistic for users to stop clicking on all links in email. Therefore, it is important to teach users how to identify links that are likely to lead to fraudulent websites and teach users not to click on those links. However, the process of identifying fraudulent links is complex.

We selected three representative tutorials from well-known sources for further evaluation: eBay's tutorial on spoofed emails [eBay 2006], Microsoft's Security tutorial on Phishing [Microsoft Corporation 2006], and Phishing E-card from the U.S. Federal Trade Commission [Federal Trade Commission 2006a]. Because none of these tutorials provide much information on parsing URLs—a skill that can help people identify fraudulent links—we also selected a URL tutorial from the online security education portal MySecureCyberspace [MySecureCyberspace 2007].

Table II presents information about the format and length of the training materials we evaluated, and summarizes the concepts taught by each. Most of the training materials we examined present a basic definition of phishing, highlight common characteristics of phishing emails, provide suggestions to avoid falling for these scams, and offer information about what to do after falling for them. The materials also provide a link to other resources about phishing and security. A common message of most of these materials was that trusted organizations will not request personal information through email.

The training materials we selected made minimal use of the basic instructional design principles introduced in Table I. The eBay and Microsoft tutorials used the contiguity principle, while the FTC video used the personalization and story-based agent environment principles. We found that some illustrations were used more for decorative purposes than for explanative purposes, and those used for explanative purposes sometimes lacked captions or explanations in the body of the text. In some cases text and associated images were located far apart from each other, either much further down on a long web page or on a different web page altogether.

Table II. Characteristics of selected existing online training materials

Source	Format	Length			Concepts taught	
		Words	Printed pages	Graphic examples	Cues to look for	Guidelines
Microsoft	Web page	737	3	2	Urging urgent action; non-personalized greeting; requesting personal information	Identify fraudulent links
eBay	Web page	1276	5	8	all the above; sender email address; links in the email; legitimate vs. fake eBay address	Never click on links in email; identify fraudulent links
FTC	Video	N/A	N/A	N/A	Requesting personal information	Never respond to emailed requests for personal information
MySecure Cyberspace	Web page	236	1	0	N/A	N/A

4.2 User study

We conducted a user study to evaluate the effectiveness of the selected online training materials. Fourteen participants were asked to examine 10 websites and determine which were phishing. They were then given 15 minutes to read the four selected training materials. After training, the participants were asked to examine 10 more websites and determine which were phishing. The phishing websites in this study were simulated on the local machine by hacking the local host file. A control group of fourteen participants completed the same protocol, but spent the 15-minute break playing solitaire and checking their email instead of reading training materials.

We measured false positives and false negatives before and after training. A false positive occurs when a legitimate website is mistakenly judged as a phishing website. A false negative occurs when a phishing website is incorrectly judged to be legitimate. We found that false negatives fell from 38% before training to 12% after training. However, false positives increased from 3% to 41%. The control group did not perform significantly differently before and after their 15-minute break. Further details of this study are discussed in Section 6.2.

The results suggest that the existing online training materials are surprisingly effective in helping users identify phishing websites when users actually read the training materials. However, they could also be made more effective by applying basic instructional design principles. Furthermore, while our results demonstrate

that users are better at avoiding phishing websites after reading the training materials, users were are also likely to have more false positives. Finally, even if more effective online training materials were available, getting users to read them voluntarily remains a problem. The rest of this paper describes the work that we have done to develop better approaches to anti-phishing training through principled instructional design, innovative delivery methods, and controlled evaluation.

5. PHISHGURU

In this section, we discuss the design and evaluation of PhishGuru,[1] an embedded training system that teaches people about phishing while in their normal use of email. In Section 5.1 we present the design of PhishGuru and describe our design rationale. In Section 5.2 we present a laboratory study measuring knowledge acquisition after PhishGuru training [Kumaraguru et al. 07a]. In Section 5.3, we present a study measuring users' knowledge retention and knowledge transfer after PhishGuru training [Kumaraguru et al. 07b].

5.1 Design of PhishGuru

Embedded training is a training method in which instructional materials are integrated into the primary tasks that learners perform in their everyday lives. This method has been widely applied in training military personnel on new Future Combat Systems (FCS) [Kirkley and et al. 2003; Burmester et al. 2005]. Education researchers believe that training is most effective when training materials incorporate real-world context [Anderson and Simon 1996; Clark and Mayer 2002].

Phishing attacks are carried out most frequently by sending victims email messages that direct them to fraudulent websites. Thus, there are two primary intervention points for an anti-phishing training system: email messages and websites. We focus on the email intervention point rather than websites for three reasons. First, email is the main vector for delivering phishing messages to users. If we can prevent people from trusting phishing emails, they will not visit phishing websites. Second, anti-phishing websites (such as those reviewed in Section 4) require learners to proactively visit them, limiting the number of people who will actually see these websites. In contrast, an embedded training approach brings information directly to learners. Third, learners must already have some knowledge about phishing or other kinds of scams to seek out educational websites. In contrast, embedded training works for both experts and non-experts who are unaware of phishing, by educating learners immediately after they have made a mistake.

In PhishGuru, users are periodically sent training emails in the form of simulated phishing emails, perhaps from their system administrator or from a training company. Users access these training emails in their inbox while they are checking their regular emails. These training emails look just like phishing emails, urging people to go to some website and login. If people fall for the training email (that is, they click on a link in that email), we provide an intervention message that explains that they are at risk for phishing attacks and give some tips to users for protecting themselves. Providing immediate feedback at this "teachable moment" enhances learning [Anderson 1993], [Mathan and Koedinger 2005].

[1] http://phishguru.org/

There is a plethora of literature on *teachable moments*, for example in the area of sexual behaviors and HIV prevention, injury prevention, and smoking cessation [McBride et al. 2003]. When users click on a link in a PhishGuru training email, they are shown a training message that alerts them to the risk of clicking on links, thereby creating a teachable moment that can influence user behavior. The embedded training approach provides training at a time when learners are likely to be motivated to learn, without requiring learners to proactively seek out training or allocate time in their schedules for training. In addition, it enables a system administrator or training company to train people continuously as new threats arise.

We created and evaluated several prototypes of our embedded training system. One early design consideration was whether to show interventions immediately after a person had clicked on a training email or only after they had tried to login to the fake website. Our pilot test with paper prototypes strongly suggested that showing an intervention immediately after a person had clicked on a link was better, since people who were shown interventions after logging in were confused as to why they were seeing warning messages about the risks of clicking on email links. We believe this is due to a time delay between cause (clicking on a link) and effect (seeing a warning message about email after logging in). Egelman et al. observed a similar gap when user study participants who had seen a web browser anti-phishing warning returned to the email that triggered the warning and repeatedly tried to access the fraudulent website, unaware that the email itself was fradulent [Egelman et al. 2007].

Informed by our early designs, we created two new interventions: a text/graphic intervention (see Figure 1) and a comic strip intervention (see Figure 2). The text/graphic intervention describes the risks of phishing, shows a small screenshot of the training email, points out cues that it is a phishing email, and outlines simple actions that users can take to protect themselves. The comic strip intervention conveys similar information as the text/graphic intervention, but in a comic strip format. Both interventions feature the five guidelines identified in Section 4. Table III summarizes the ways in which we applied the instructional design principles discussed in Section 3.1 to the design of PhishGuru.

To test the effectiveness of PhishGuru we measured knowledge acquisition (Section 5.2), as well as knowledge retention and transfer (Section 5.3).

5.2 Evaluation of Knowledge Acquisition

5.2.1 *Study design.* This study was designed to measure knowledge acquisition after PhishGuru training. As our research was focused on novice users, we recruited participants with little technical knowledge. We posted fliers around our university and local neighborhoods, and then screened users through an online survey. We recruited 30 participants who said they had done no more than one of the following: changed preferences or settings in their web browser, created a web page, and helped someone fix a computer problem. Each participant was randomly placed in one of three conditions: "notices," "text/graphic," and "comic." Participants in the "notices" condition were shown typical security notices in the form of an email in which users are requested to go to the organization's website by clicking on a link if they are interested in learning how to identify spoof emails. Participants

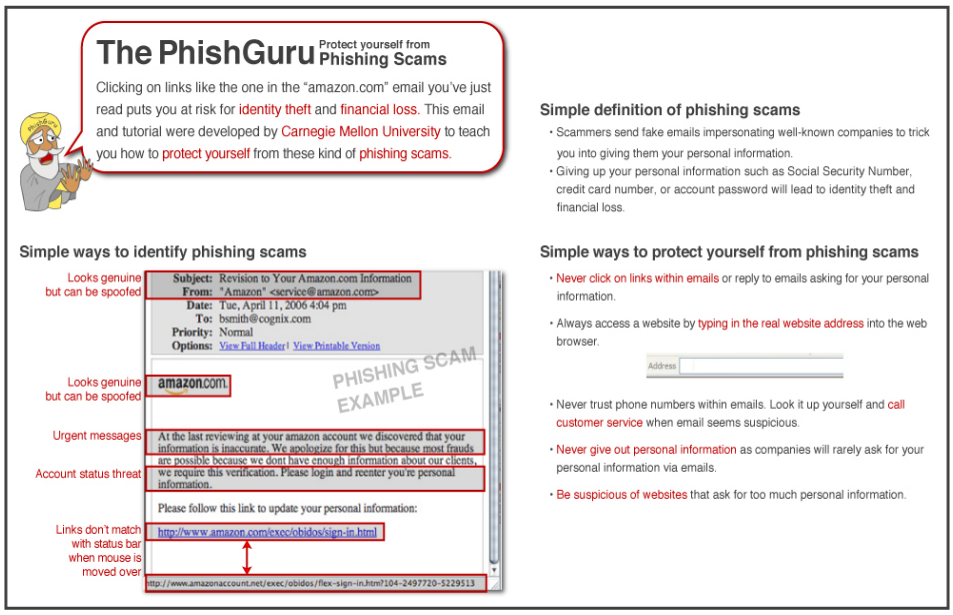

Fig. 1. The text/graphic intervention design used in the knowledge acquisition study shows an annotated image of the training email that led to this warning.

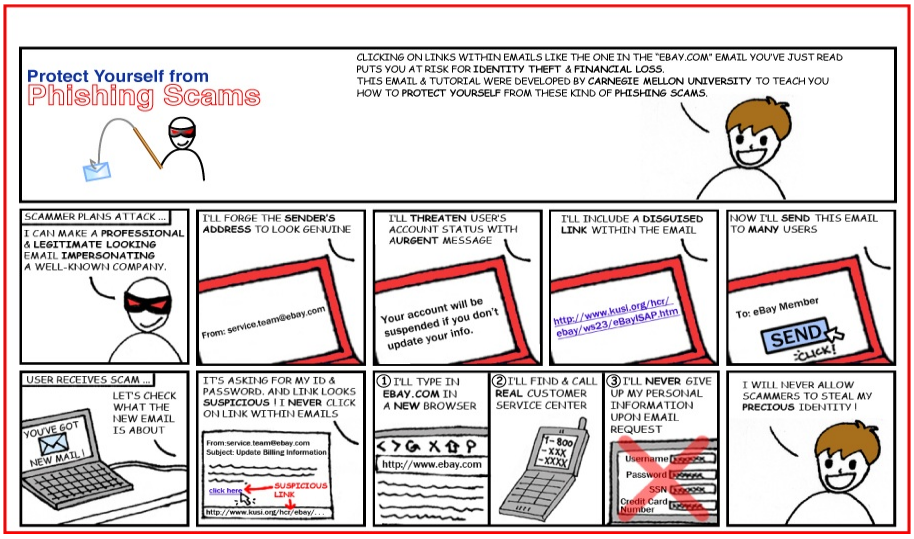

Fig. 2. The comic strip intervention design used in the knowledge acquisition study presents information using the design principles discussed in Section 3.1.

Table III. Application of instructional design principles to the design of PhishGuru

Principle	**Way(s) in which we applied the principle in our design**
Learning-by-doing	The training materials are presented when users fall for phishing emails. Thus, users learn by actually clicking on phishing emails (doing).
Immediate feedback	We provide feedback through interventions immediately after the user clicks on a link in a simulated phishing email.
Conceptual-procedural	Our interventions present a definition of phishing (conceptual knowledge) side-by-side with ways to avoid falling for phishing attacks (procedural knowledge)
Contiguity	We have placed pictures and relevant text contiguously in our interventions, for example instructions 1 through 3 in Figure 2.
Personalization	We use a conversational style throughout our interventions, using the words, "I" and "you."
Story-based agent environment	The comic strip intervention tells a story using two characters, a phisher and a victim.

in the "text/graphics" condition were shown the text and graphics intervention. Participants in the "comic" condition were shown the comic strip intervention.

The user study consisted of a think-aloud session in which participants played the role of "Bobby Smith," an employee of Cognix Inc. who works in the marketing department. Participants were told that the study investigated "how people effectively manage and use emails." They were told that they should interact with Bobby Smith's email the way they would normally interact with their own email. We handed them a list of Bobby's accounts with usernames and passwords. We used a completely functional SquirrelMail[2] implementation for users to access Bobby Smith's email. We wrote a Perl script to push emails into the SquirrelMail server and used this script to change the training emails for each group. We designed the emails in Bobby's inbox to allow us to measure the effectiveness of our interventions.

Each participant was shown 19 email messages, arranged in a predefined order, shown in Table IV. Nine *legitimate-no-link* messages were legitimate emails without any links, received from co-workers at Cognix, friends, and family. These emails requested that Bobby Smith perform simple tasks such as replying. Two *legitimate-link* messages were simulated legitimate emails from organizations with which Bobby Smith had an account. The mailbox contained two *spam* emails, two *phishing-account* fraudulent emails that appeared to come from organizations where Bobby has an account, and two *phishing-no-account* fraudulent emails that appeared to come from a bank with which Bobby does not have an account. The mailbox also had two *training* emails—security notices or embedded training interventions. Table V presents examples of each type of email that we used in the study.

Table VI gives the demographic details of the study participants in both the knowledge acquisition (KA) study described here and the knowledge retention and knowledge transfer (KR and KT) study described in Section 5.3. We found no cor-

[2]http://www.squirrelmail.org/

relation between participant demographics and susceptibility to falling for phishing emails in our study.

Table IV. Email arrangement in the study

1. Legitimate-no-link	11. Training
2. Legitimate-no-link	12. Spam-link-no-account
3. Phishing-account	13. Legitimate-link-no-account
4. Legitimate-no-link	14. Phishing-no-account
5. Training	15. Legitimate-no-link
6. Legitimate-no-link	16. Phishing-no-account
7. Legitimate-link-account	17. Phishing-account
8. Spam-link-no-account	18. Legitimate-no-link
9. Legitimate-no-link	19. Legitimate-no-link
10.Legitimate-no-link	

Table V. Sample of emails used in the PhishGuru study

Email type	**Sender information**	**Email subject line**
Legitimate-no-link	Brandy Anderson	Booking hotel rooms for visitors
Legitimate-link	Josept Dicosta	To check the status of the product on Staples
Phishing-no-account	Wells Fargo	Update your bank account information!
Phishing-account	PayPal	Reactivate you PayPal account!
Spam	Eddie Arredondo	Fw: Re: You will want this job
Training	Amazon	Revision to your Amazon.com information

5.2.2 *Results.* In this study we measured knowledge acquisition based on whether users clicked on links in legitimate-link, phishing-no-account, and phishing-account emails before and after training. We found significant differences in knowledge acquisition between the three study conditions, with participants in the comic condition demonstrating the greatest knowledge acquisition. Participants in the comic group were significantly better at recognizing phishing emails than those in the notices group (Chi-square test, $p < 0.01$). Participants in the text/graphics group also performed better than those in the notices group, but this difference was not significant. Participants in the comic condition spent the most time reading training materials. Average time spent reading training materials was 8.5 seconds (SD = 11) for the notice condition, 107 seconds (SD = 21) for the text/graphics condition, and 120 seconds (SD = 24) for the comic condition. Figure 3 presents a comparison of the three training methodologies for all the emails that had links in them.

In the security notice condition nine participants (90%) clicked on links in phishing-account emails before the security notice messages and the same number of participants clicked on phishing-account links after the security notice messages. The participants who viewed the security notices said that the information took too long

Table VI. Participants in the PhishGuru studies

Characteristics	KA Study			KR and KT Study			
	Notices	Text/ Graphic	Comic	Control	Suspicion	Embedded	Non-Embedded
Sample size	10	10	10	14	14	14	14
Gender							
Male	50%	40%	20%	36%	50%	36%	43%
Female	50%	60%	80%	64%	50%	64%	57%
Browser							
IE	80%	60%	60%	64%	36%	50%	50%
Firefox	10%	20%	30%	29%	57%	29%	43%
Others	10%	20%	10%	7%	7%	21%	7%
Average emails per day	51.4	36.9	15	20.5	17.6	16.1	20.7
Average age in years	31.2	27.5	21.1	28	26.9	24.6	24.3

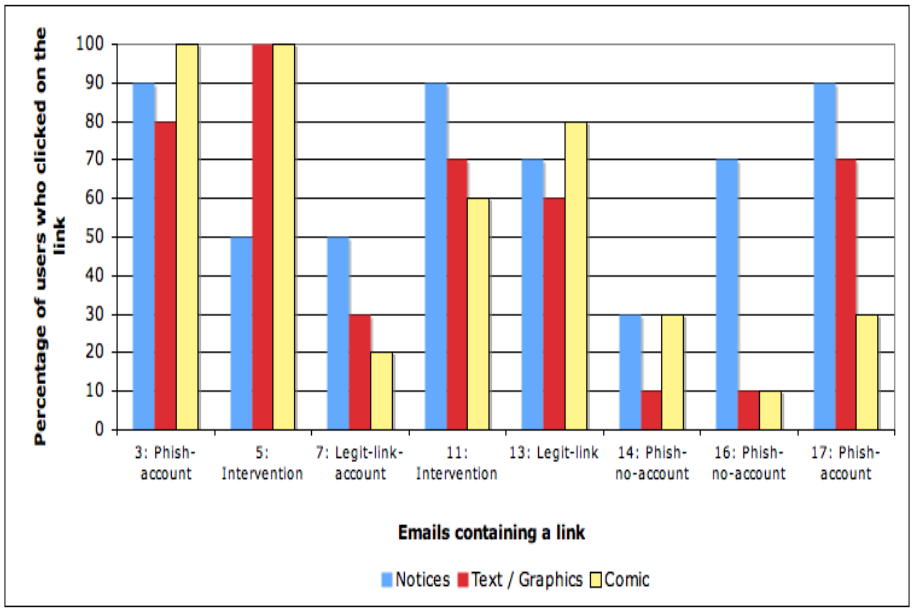

Fig. 3. Percentage of users who clicked on the links in each condition

to read and they were not sure what the messages were trying to convey. The mean percentage of participants falling for the three phishing emails presented after the security notices was 63%.

In the text/graphic condition, 80% of the participants fell for the first phishing-account email (before any training) while all participants clicked on the training message link in the training email. Seven participants (70%) clicked on the second

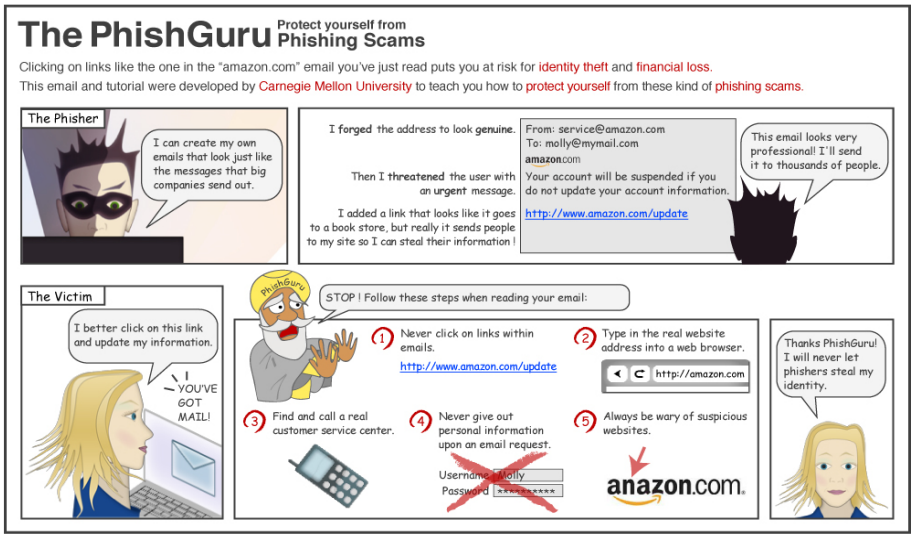

Fig. 4. Revised comic strip intervention design used in the knowledge retention and transfer study. The top row presents the activities of a phisher while the bottom row shows the victim and presents steps the victim can take to avoid falling for the phishing attack.

training message and seven participants (70%) fell for the final phishing-account email. The mean percentage of participants falling for the three phishing emails presented after the interventions was 30%.

In the comic strip condition, all participants fell for the first phishing email and also clicked on the training message. Six participants (60%) clicked on the second training message and only three participants (30%) fell for the final phishing email. The mean percentage of participants falling for the three phishing emails presented after our interventions was 23%.

Our results suggest that the current practice of sending out security notices is ineffective. Our results also indicate that both of our embedded training interventions helped teach people how to avoid phishing attacks. Our comic strip intervention—which has significantly less text and more graphics than our text/graphics intervention and tells a story to convey its message—was the most effective intervention. This study also demonstrated that when users fell for our simulated phishing attacks they were motivated to spend time reading training materials.

5.3 Evaluation of Knowledge Retention and Knowledge Transfer

In Section 5.2 we evaluated PhishGuru by testing users immediately after training. In this section, we present the results of a study looking at (1) how well PhishGuru users can retain and transfer knowledge, and (2) how important it is to fall for simulated phishing attacks, as opposed to simply getting PhishGuru interventions directly as email messages [Kumaraguru et al. 07b].

5.3.1 *Study design.* Using the lessons learned from the study discussed in Section 5.2, we created a new version of the comic strip intervention, as shown in Figure 4. We added "The PhishGuru" as a character in the revised design, reduced

the amount of text, and made the instructions clearer.

Our second study used a similar design as the previous study, once again asking participants to play the role of Bobby Smith. We had four conditions: "embedded," "non-embedded," "suspicion" and "control." Participants in the embedded condition received a simulated phishing email and saw the revised comic strip intervention 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, without having to fall for a simulated phishing email. Participants in the suspicion condition received a brief email from a friend that mentioned phishing, without providing any information about how they could protect themselves. Participants in the control condition received an additional email from a friend, but received no training.

The recruitment and the lab setup were exactly the same as the study described in Section 5.2. This study was done in two parts (seven days apart). In the first part, participants saw 33 emails in Bobby's inbox: a set of 16 emails, a training intervention, and a set of 16 additional emails shown immediately after training. In the second part, the participants saw another 16 emails in Bobby's inbox. Of the 16 emails before and after the training, there were 9 legitimate-no-link emails, 3 legitimate-link emails, 2 phishing-account emails, 1 phishing-no-account email, and 1 spam email. The emails used in this study was also similar to the samples discussed in Table V. Participant demographics are shown in Table VI.

Hypotheses: In this study we tested the following three hypotheses:

(1) Participants in the embedded condition learn more effectively than participants in the non-embedded condition, suspicion condition, and the control condition.

(2) Participants in the embedded condition retain more knowledge about how to avoid phishing attacks than participants in the non-embedded condition, suspicion condition, and the control condition.

(3) Participants in the embedded condition transfer more knowledge about how to avoid phishing attacks than participants in the non-embedded condition, suspicion condition, and the control condition.

5.3.2 *Results.* Our results support our three hypotheses, providing evidence that embedded training enhances knowledge acquisition, knowledge retention, and knowledge transfer, allowing learners to effectively identify phishing messages without misidentifying legitimate messages.

To test Hypothesis 1, we measured the percentage of correct decisions that participants in each condition made for phishing and legitimate-link emails before and after the training. Our results support Hypothesis 1, demonstrating that participants in the embedded training condition learned to detect phishing-account emails effectively while participants in the other conditions did not. Participants in the embedded and non-embedded conditions did not perform significantly differently in correctly identifying phishing-account emails before the training (two sample t-test, $p = 0.19$). However, those in the embedded condition performed significantly better than those in the non-embedded condition immediately after training (two

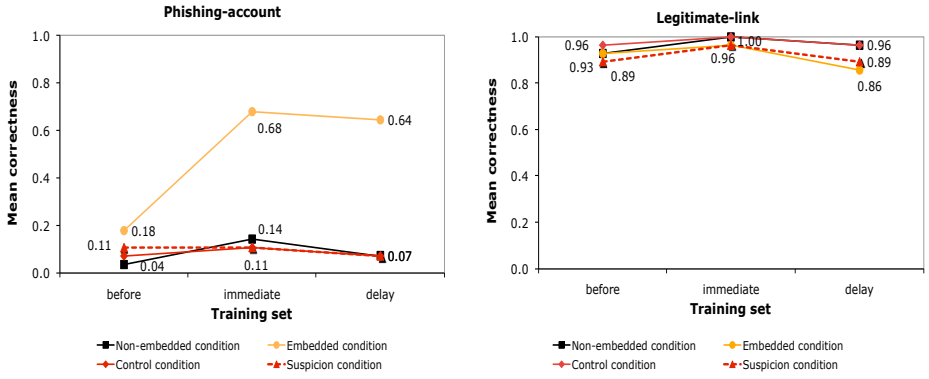

Fig. 5. Mean correctness for identifying phishing-account and legitimate-link emails before training, immediately after training, and after a one-week delay.

sample t-test, $p < 0.01$), as shown in the left side of Figure 5. Those in the embedded condition improved their performance significantly immediately after the training (paired t-test, $p < 0.01$), while those in the non-embedded condition did not (paired t-test, $p = 0.27$). Those in the embedded condition also spent more time reading the intervention. On average, participants in the embedded condition spent 97 seconds (SD = 66), while participants in the non-emebedded condition spent 37 second (SD = 32).

To test Hypothesis 2 and measure knowledge retention after a one-week delay, we compared performance on phishing-account and legitimate-link emails before, immediately after training, and after a one-week delay. Our results support Hypothesis 2 and suggest that users in the embedded condition were able to retain the knowledge they acquired and use it to distinguish phishing and legitimate emails, even after a one-week delay. In all conditions there was no significant difference between mean correctness on phishing-account emails or legitimate-link emails immediately after the training and after a one-week delay. Participants in the embedded condition improved their performance significantly on phishing-account emails after the delay compared to before the training (paired t-test, $p = 0.02$), while participants in the non-embedded group did not improve (paired t-test, $p = 0.67$), as shown in the left side of Figure 5. While 64% of the embedded-condition participants identified the phishing-account email correctly after a one-week delay, only 7% of the participants in the other conditions identified the email correctly.

The phishing-account email sent immediately after training simulated an account update request that appeared to come from Citibank, similar to the training message. To measure knowledge transfer, we used a phishing-account email 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, $p < 0.01$). 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.

Our results reinforce and extend the findings of our previous study, which sug-

gested that embedded training can be an effective method to train users to distinguish between legitimate and phishing email messages. The fact that we saw no significant performance drop-off after one week suggests that users are likely to retain their training for longer time periods. In the delay state, we found significant differences between the non-embedded and the embedded training conditions in correctly identifying the eBay email (phishing-account type) as a phishing attack. This shows that participants in the embedded condition were able to transfer knowledge to another situation better than participants in the non-embedded condition. Our observation that the suspicion condition was not significantly different from the control condition suggests that it is not helpful to tell users about phishing without providing them with information about how to identify phishing or actionable steps they can take to protect themselves. Finally, our observation that the non-embedded condition was not significantly different from the control condition provides a strong indicator that the delivery method is an important factor in determining the effectiveness of anti-phishing training. Indeed we found that our revised comic strip intervention, which provided fairly effective training when displayed at that teachable moment after participants had fallen for a simulated phishing attack, was completely ineffective when sent directly via email.

6. ANTI-PHISHING PHIL

In this section we discuss Anti-Phishing Phil,[3] an educational game we designed to train users about phishing attacks. Anti-Phishing Phil motivates users to learn by embedding training into a fun activity. In addition, the highly interactive nature of the game allows us to teach users how to distinguish legitimate links from fraudulent ones and provide users with immediate opportunities to practice this procedure multiple times. Thus, Anti-Phishing Phil complements PhishGuru by providing an entertaining platform for the rapid repetition and feedback needed to teach more difficult anti-phishing procedures.

In Section 6.1 we present the design of Anti-Phishing Phil and describe the ways in which we applied instructional design principles in designing the game. In Section 6.2, we present a laboratory study evaluation [Sheng et al. 2007]. In Section 6.3 we present new results from a field study.

6.1 Design of Anti-Phishing Phil

The main character of the game is a young fish named Phil. 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 websites) 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. He helps Phil out by providing tips on how to identify bad worms (and hence, phishing websites).

The game is split into four rounds, each of which is two minutes long. Before each round begins, users must view a short tutorial that provides anti-phishing tips, as shown in Figure 6. In each round, Phil is presented with eight worms, each

[3]http://cups.cs.cmu.edu/antiphishing_phil/

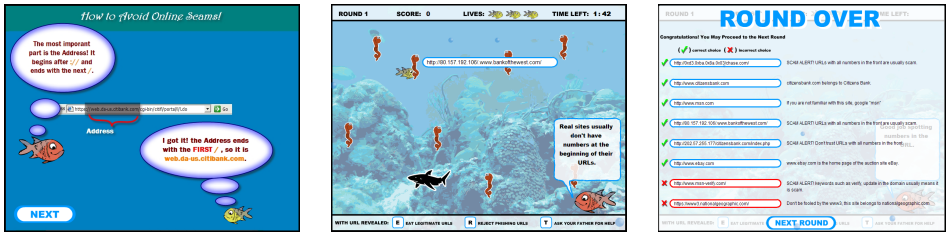

Fig. 6. Screen shots from Anti-Phishing Phil. The left screen shows part of one of the tutorials shown before the beginning of a round. The middle screen shows a URL being displayed as Phil swims by a worm and a tip from the father fish. The right screen shows the end of round summary.

of which carries a URL that is displayed when Phil moves near it, as shown in Figure 6. The player can move Phil around the screen and "eat" the real worms or "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. 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. At the end of every round a review screen is displayed, showing all of the URLs from that round and tips for identifying them correctly, as shown in Figure 6.

The game is implemented in Flash 8. The content for the game, including URLs and training messages, is loaded from a separate data file at the start of the game. This 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 use sound effects to provide audio feedback, and background music and underwater background scenes to help keep users engaged.

We used the *educational action design* methodology to design the game. In this method, the learner is given a stipulated time in which they have to perform (and thereby learn) the things that are presented in the game [Baker et al. 2007]. Table VII summarizes the ways in which we applied instructional design principles in designing Anti-Phishing Phil.

6.2 Anti-Phishing Phil lab study

6.2.1 *Study design.* We conducted a user study using the protocol introduced in Section 4.2 to measure knowledge acquisition from playing Anti-Phishing Phil. Participants were asked to examine 10 websites and determine which were phishing. After 15 minutes of training they were asked to examine 10 more websites and determine which were phishing. Half of the websites were phishing websites from popular brands and half were legitimate websites from popular financial institutions and online merchants, as well as random websites.

As this research is also focused on educating novice users about phishing attacks, we recruited participants with little technical knowledge. We posted fliers around our university and local neighborhoods, and then screened users through an online

Table VII. Applying the instructional design principles in Phil design

Principle	**Ways in which we applied the principle in our design**
Learning-by-doing	Users identify real and fake websites while playing a game
Immediate feedback	We provide feedback through points, lives, and end of round summary
Conceptual-procedural	Applied in the between-round tutorials, for example, we provide 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 websites (conceptual knowledge)
Contiguity	Applied in the between-round tutorials
Personalization	Applied in the messages from the father fish
Story-based agent environment	Applied by having the user control a young fish named Phil (agent), who has to learn anti-phishing skills to survive in water among sharks and big fishes (story)
Reflection	Applied at the end of each round by displaying a list of websites that appeared in that round and an indication as to whether the user correctly or incorrectly identified each one

survey. We recruited 28 participants and assigned them randomly to either a "tutorial" condition or "game" condition. In the tutorial condition, participants were asked to spend up to fifteen minutes reading an anti-phishing tutorial we created based on the Anti-Phishing Phil game. The tutorial included 17 pages of color printouts of all of the between-round training messages and 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. In the game condition, participants played the Anti-Phishing Phil game for fifteen minutes.

We compare the results of the Anti-Phishing Phil lab study with the data from the existing training material evaluation presented in Section 4.2. Table VIII shows the demographic details of the participants in both studies.

6.2.2 *Results.* We measured learners' knowledge acquisition from playing Anti-Phishing Phil by examining false positives, false negatives, and the total percentage of correct websites identified before and after playing the game. A false positive occurs when a legitimate website is mistakenly judged as a phishing website. A false negative occurs when a phishing website is incorrectly judged to be a legitimate website. As shown in Figure 7, 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, but this was not statistically significant.

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, $p = 0.01$); for the tutorial condition, it changed from 0.43 to 0.19 (paired t-test, $p < 0.03$); for the game condition, it changed from 0.34 to 0.17 (paired t-test, $p < 0.02$). There is no statistical difference between the groups in either the pre-test (oneway ANOVA, $p = 0.60$), or post-test (oneway ANOVA, $p = 0.45$). Post-test false positive rates decreased significantly in the game condition ($p < 0.03$). A one-way ANOVA revealed that false positive rates

Table VIII. Participants for the Anti-Phishing Phil study.

Characteristics	Conditions			
	Existing training material	Tutorial	Game	Control
Sample size	14	14	14	14
Gender				
Male	29%	36%	50%	33%
Female	71%	64%	50%	67%
Age				
18 - 34	93%	100%	100%	93%
> 34	7%	0%	0%	7%
Education				
High School	14%	7%	7%	9%
College Undergrad	51%	79%	51%	48%
College graduate	14%	7%	21%	22%
Post. Graduate school	21%	7%	21%	22%
Years on the Internet				
3- 5 years	23%	23%	15%	15%
6-10 years	69%	70%	78%	70%
> 11 years	8%	7%	7%	15%

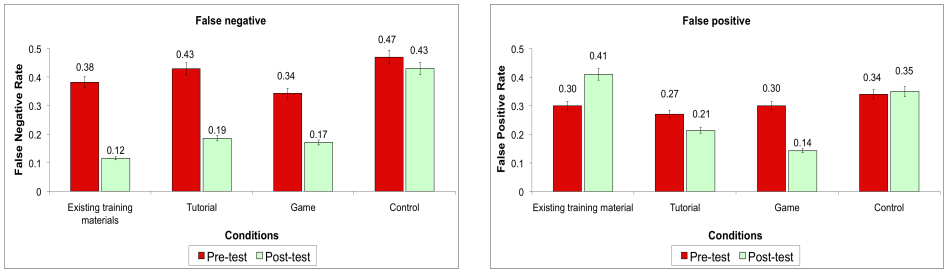

Fig. 7. False negatives and false positives on pre-test and post-test. The differences in false negatives between groups are not statistically significant. The game condition has significantly lower false positives than the existing training materials.

differed significantly in the post-test (paired t-test, $p < 0.02$). 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.

Our results demonstrate that users show significant improvements in their ability to identify phishing links correctly after 15 minutes of training with Anti-Phishing Phil, our tutorial, or existing online training materials. However, participants in the game condition were better able to distinguish between phishing and legitimate links than those in the other conditions, and were thus less likely to incorrectly identify legitimate links as phishing links.

6.3 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.

6.3.1 *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. In addition, several press reports about Anti-Phishing Phil directed people to our study website. 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.

6.3.2 *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.

6.3.3 *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). We had 46 participants in the novice group, 256 in intermediate and 372 in expert. As illustrated in Figure 8, novice users showed the greatest improvement, with false positive rate decreasing from 0.84 to 0.22 (paired t-test, $p < 0.0001$), and false negative rate decreasing from 0.57 to 0.22 (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 website'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

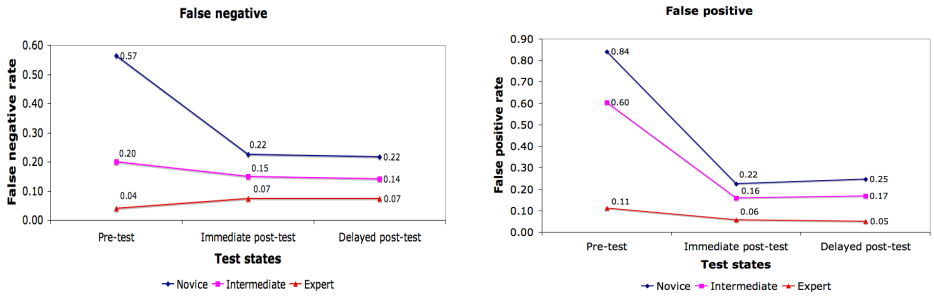

Fig. 8. False negative and false positive for Anti-Phishing Phil in the real-world. Novice users show greatest improvement in false negative and false positive.

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. A Chi-square test of the percentage of novice, intermediate and expert users who completed the immediate post-test, or delayed post-test 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 Tukey 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 were 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 ($r = -0.645$, $p < 0.001$). We found that users are most confused by long URLs, making 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).

7. 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

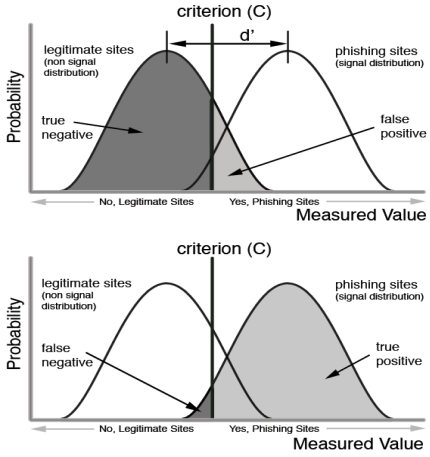

Fig. 9. Applying signal detection theory (SDT) to anti-phishing education We treat legitimate websites as "non signal," and phishing websites as "signal." Sensitivity (d') measures users' ability to distinguish signal from non-signal. Criterion (C) measures users' decision tendency ($C < 0$ indicates 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.

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) [Salkind 2006; Macmillan and Creelman 2004] 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 9 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 cautious, 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 our evaluation of existing online training materials, PhishGuru retention and transfer study, Anti-Phishing Phil laboratory study, and

Anti-Phishing Phil field study[4]. Table IX summarizes the results from the analysis. 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 IX. 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). A two-sample t test was performed pre- and post-test.

	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*	–
PhishGuru knowledge retention and transfer (embedded condition)	0.54	2.27	1.43	1.19	0.67	0.35
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

8. CONCLUSIONS

We analyzed the learning science literature to select the principles which are most powerful and most widely used in developing educational materials. We used these learning science principles to analyze the existing online training materials. We found that existing online training materials we selected made minimal use of the basic instructional design principles. We developed and evaluated two approaches to anti-phishing user education: PhishGuru, a system to educate people about phishing during their normal use of email, and Anti-Phishing Phil, a game that teaches people how to identify phishing URLs.

Results from our PhishGuru studies suggest that the current practice of sending out security notices is ineffective, but embedded training can effectively teach people how to avoid phishing attacks. Results from our Anti-Phishing Phil studies

[4]http://www.aston.ac.uk/downloads/lhs/georgema/d'_calculator4.xls

demonstrate that participants who played the game performed better at identifying phishing websites than participants who completed two other types of training. In our evaluation of both approaches we found that people could retain what they learned for at least one week without significant degradation in performance. We believe these approaches can be used to train users about a variety of cyber security threats.

ACKNOWLEDGMENTS

The authors would like to thank all members of the Supporting Trust Decisions project for their feedback. The authors would also like to thank Dr.Vincent Aleven, Sharique Hasan, Elizabeth Nunge, and Yong Rhee for their help in conducting users studies and for fruitful discussions; and Bryant Magnien and Patrick Kelley for Anti-Phishing Phil graphics and Flash development.

REFERENCES

Abu-Nimeh, S., Nappa, D., Wang, X., and Nair, S. 2007. A comparison of machine learning techniques for phishing detection. *e-Crime Researchers Summit, Anti-Phishing Working Group*.

Account Guard. 2006. Account Guard. Retrieved Nov 3, 2006, http://pages.ebay.com/ebay_toolbar/.

Adams, A. and Sasse, M. A. 1999. Users are not the enemy. *In the Communications of the ACM 42,* 12, 40–46. DOI=http://doi.acm.org/10.1145/322796.322806.

Aleven, V. and Koedinger, K. R. 2002. An effective metacognitive strategy: learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science 26,* 2, 147 – 179.

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

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

Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. 1995. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences 4,* 2, 167–207.

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

Anti-Phishing Working Group. 2007. Anti-Phishing Working Group. Retrieved Jan 9, 2007, http://www.antiphishing.org/.

Bahrick, H. P. 1979. Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology 108,* 3 (September), 296–308.

Baker, R., Habgood, J., and Ainsworth, S. E. 2007. Modeling the acquistion of fluent skill in educational action games. *Proceedings of User Modeling*, 17 – 26.

Barnett, S. M. and Ceci, S. J. 2002. When and where do we apply what we learn? a taxonomy for far transfer. In *Psychological Bulletin*. Vol. 128. 612–637.

Bransford, J. D. and Schwartz, D. L. 2001. Rethinking transfer: A simple proposal with multiple implications. In *Review of Research in Education*, A. Iran-Nejad and P. D. Pearson., Eds. Vol. 24. American Educational Research Association (AERA) Washington, DC, 61 – 100.

Burmester, G. M., Stottler, D., and Hart, J. L. 2005. Embedded training intelligent tutoring systems (ITS) for the future combat systems (FCS) command and control (C2) vehicle. Tech. rep., Defense Technical Information Center. http://www.stottlerhenke.com/papers/IITSEC-02-ITSFCS.pdf.

Chandrasekaran, M., Narayanan, K., and Upadhyaya, S. 2006. Phishing email detection based on structural properties. *NYS Cyber Security Conference*.

Clark, R. C. 1989. *Developing Technical Training: A Structured Approach for the Development of Classroom and Computer-Based Instructional Materials*. Addison Wesley Publishing Company, Beverly, MA, USA.

Clark, R. C. and Mayer, R. E. 2002. *E-Learning and the science of instruction: proven guidelines for consumers and designers of multimedia learning.* John Wiley & Sons, Inc., USA.

Committee on Developments in the Science of Learning and National Research Council. 2000. *How People Learn: Bridging Research and Practice.* National Academies Press.

Corbett, A. T. and Anderson, J. R. 2001. Locus of feedback control in computer-based tutoring: impact on learning rate, achievement and attitudes. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM Press, New York, NY, USA, 245–252.

Cordova, D. I. and Lepper, M. R. 1996. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology 88,* 4 (December), 715–730.

Cranor, L. F. 2008. A Framework for Reasoning About the Human In the Loop. *Usability, Psychology and Security.*

Cranor, L. F. and Garfinkel, S. Aug, 2005. *Security and Usability: Designing Secure Systems that People Can Use.* O'Reilly, Sebastopol, CA, USA.

Dhamija, R. and Tygar, J. 2005. The Battle Against Phishing: Dynamic Security Skins. In *SOUPS '05: Proceedings of the 2005 symposium on Usable privacy and security.* ACM Press, New York, NY, New York, NY, USA, 77–88. Retrieved Feb 10, 2006, DOI= http://doi.acm.org/10.1145/1073001.1073009.

eBay. 2006. Spoof email tutorial. Retrieved December 30, 2006. http://pages.ebay.com/education/spooftutorial.

Eberts, R. E. 1997. *Handbook of Human-computer Interaction.* Elsevier Science, Chapter Computer-based Instruction, 825–847.

Egelman, S., Cranor, L. F., and Hong, J. 2007. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. To appear in CHI '08: Proceedings of the SIGCHI conference on Human factors in computing systems.

Emigh, A. 2005. Online identity theft: Phishing technology, chokepoints and countermeasures. Tech. rep., Radix Labs. October. http://www.antiphishing.org/Phishing-dhs-report.pdf.

Evers, J. 2006. User education is pointless. http://news.com.com/2100-7350_3-6125213.html.

Federal Trade Commission. 2006a. An e-card for you game. Retrieved December 30, 2006. http://www.ftc.gov/bcp/conline/ecards/phishing/index.html.

Federal Trade Commission. 2006b. How not to get hooked by a phishing scam. Consumer alert news. Retrieved Nov 7, 2006, http://www.ftc.gov/bcp/edu/pubs/consumer/alerts/alt127.htm.

Ferguson, A. J. 2005. Fostering E-Mail Security Awareness: The West Point Carronade. *EDUCASE Quarterly* 1. http://www.educause.edu/ir/library/pdf/eqm0517.pdf.

Fette, I., Sadeh, N., and Tomasic, A. 2006. Learning to detect phishing emails. *16th International conference on World Wide Web.*

Florencio, D. and Herley, C. 2005. Stopping a phishing attack, even when the victims ignore warnings. Tech. rep., Microsoft.

Fong, G. T. and Nisbett, R. E. 1991. Immediate and delayed transfer of training effects in statistical reasoning. In *American Psychological Association Inc.* Vol. 120. Journal of Experimental Psychology, 34–45.

Gagne, R. M., Foster, H., and Crowley, M. E. 1948. The measurement of transfer of training. *Psychological Bulletin 45,* 2, 97–130.

Gordon, L. A., Loeb, M. P., Lucyshyn, W., and Richardson, R. 2006. CSI/FBI Computer Crime and Security Survey. Report, Computer Security Institute.

Gorling, S. 2006. The myth of user education. In *Proceedings of the 16th Virus Bulletin International Conference.*

Hight, S. D. 2005. The importance of a security, education, training and awareness program. http://www.infosecwriters.com/text_resources/pdf/SETA_SHight.pdf.

Jackson, C., Simon, D., Tan, D., and Barth, A. 2007. An evaluation of extended validation and picture-in-picture phishing attacks. In *Usable Security (USEC'07).* http://usablesecurity.org/papers/jackson.pdf.

Jagatic, T., Johnson, N., Jakobsson, M., and Menczer, F. 2007. Social phishing. *In the Communications of the ACM 50,* 10 (October), 94–100. Retrieved March 7, 2006, http://www.indiana.edu/ phishing/social-network-experiment/phishing-preprint.pdf.

Jakobsson, M. 2007. The human factor in phishing. In *Privacy & Security of Consumer Information*. http://www.informatics.indiana.edu/markus/papers/aci.pdf.

Jakobsson, M. and Myers, S., Eds. 2006. *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*. Wiley-Interscience.

James, L. 2005. *Phishing Exposed*. Syngress Publishing, Canada.

Johnson, B. R. and Koedinger, K. R. 2002. Comparing instructional strategies for integrating conceptual and procedural knowledge. In *Proceedings of the Annual Meeting [of the] North American Chapter of the International Group for the Psychology of Mathematics Education*. Vol. 1–4. 969–978.

Kirkley, J. R. and et al. 2003. 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)*. http://www.iforces.org/downloads/problem-based.pdf.

Koedinger, K. R. 2002. Toward evidence for instruction design principles: Examples from cognitive tutor math 6. *Proocedings of the Annual Meeting, Norh American Chapter of the International Group for the Psychology of Mathematics Education 1 – 4*.

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., and Nunge, E. 07a. Protecting people from phishing: the design and evaluation of an embedded training email system. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press, New York, NY, USA, 905–914.

Kumaraguru, P., Rhee, Y., Sheng, S., Hasan, S., Acquisti, A., Cranor, L. F., and Hong, J. 07b. Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. *e-Crime Researchers Summit, Anti-Phishing Working Group*.

Lininger, R. and Vines, R. D. 2005. *Phishing: Cutting the Identity Theft Line*. Indianapolis, Indiana, USA.

Macmillan, N. A. and Creelman, C. D. 2004. *Detection Theory: A User's Guide*. Lawrence Erlbaum.

Mail Frontier. 2006. Mailfrontier phishing IQ test. Retrieved Sept 2, 2006, http://survey.mailfrontier.com/survey/quiztest.html.

Mandl, H. and Levin, J. R. 1989. *Knowledge Acquisition from Text and Pictures*. North - Holland.

Mathan, S. A. and Koedinger, K. R. 2003. *Artificial Intelligence in Education: Shaping the Future of Learning Through Intelligent Technolgis*. IOS Press, Chapter Recasting the Feedback Debate: Benefits of Tutoring Error Detection and Correction Skills, 13–20.

Mathan, S. A. and Koedinger, K. R. 2005. Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational Psychologist 40,* 4, 257–265.

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

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

McBride, C. M., Emmons, K. M., and Lipkus, I. M. 2003. Understanding the potential of teachable moments: the case of smoking cessation. *Health Education Research 18,* 2, 156 – 170.

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

Microsoft Corporation. 2006. Consumer awareness page on phishing. Retrieved Sep 10, 2006. http://www.microsoft.com/athome/security/email/phishing.mspx.

Miller, R. C. and Wu, M. 2005. Fighting Phishing at the User Interface. *O'Reilly*. In Lorrie Cranor and Simson Garfinkel (Eds.) Security and Usability: Designing Secure Systems that People Can Use.

Moreno, R. and Mayer, R. E. 1999. Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology 91*, 358–368.

MySecureCyberspace. 2007. Uniform resource locator (URL). Retrieved Feb 4, 2007, http://www.mysecurecyberspace.com/encyclopedia/index/uniform-resource-locator-url-.html.

Netcraft. 2006. Netcraf. Retrieved Nov 3, 2006, http://toolbar.netcraft.com/.

New York State Office of Cyber Security & Critical Infrastructure Coordination. 2005. Gone phishing... a briefing on the anti-phishing exercise initiative for new york state government. Aggregate Exercise Results for public release.

Nielsen, J. 2004. User education is not the answer to security problems. http://www.useit.com/alertbox/20041025.html.

Robila, S. A. and Ragucci, J. W. 2006. Don't be a phish: steps in user education. In *ITICSE '06: Proceedings of the 11th annual SIGCSE conference on Innovation and technology in computer science education*. ACM Press, New York, NY, USA, 237–241. DOI=http://doi.acm.org/10.1145/1140124.1140187.

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

Salkind, N. J. 2006. *Encyclopedia of Measurement and Statistics*. Sage Publications.

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

Schneier, B. 2000. Semantic attacks: The third wave of network attacks. Crypto-Gram Newsletter. Retrieved Sept 2, 2006, http://www.schneier.com/crypto-gram-0010.html#1.

Schwartz, D. L. and Bransford, J. D. 1998. A time for telling. In *Cognition & Instruction*. Vol. 16. 475–522.

Sender Policy Framework. 2006. Sender Policy Framework. Retrieved Jan 21, 2007, http://www.openspf.org/.

Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L. F., Hong, J., and Nunge, E. 2007. Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. Symposium On Usable Privacy and Security.

Sheng, S., Wardman, B., Warner, G., Cranor, L., Hong, J., and Zhang, C. 2009. An empirical analysis of phishing blacklists. *Sixth Conference on Email and Anti-Spam*.

Singley, M. and Anderson, J. R. 1989. *The Transfer of Cognitive Skill*. Harvard University Press, USA.

SpoofGuard. 2006. Spoofguard. Retrieved Sept 2, 2006, http://crypto.stanford.edu/SpoofGuard/.

SpoofStick. 2006. Spoofstick. Retrieved Sept 2, 2006, http://www.spoofstick.com/.

Whitten, A. 2004. Making security usable. Ph.D. thesis, Carnegie Mellon University.

Wu, M., Miller, R. C., and Garfinkel, S. L. 2006. Do Security Toolbars Actually Prevent Phishing Attacks? In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems. *Conference on Human Factors in Computing Systems (CHI)*, 601–610.

Yahoo. 2007. DomainKeys: Proving and Protecting Email Sender Identity. Retrieved Jan 21, 2007, http://antispam.yahoo.com/domainkeys.

Ye, Z. E. and Smith, S. 2002. Trusted paths for browsers. In *Proceedings of the 11th USENIX Security Symposium*. USENIX Association, Berkeley, CA, USA, 263–279.

Zhang, Y., Egelman, S., Cranor, L., and Hong, J. 2007. Phinding phish: Evaluating anti-phishing tools. In *14th Annual Network and Distributed System Security Symposium*. http://lorrie.cranor.org/pubs/ndss-phish-tools-final.pdf.

...