ABSTRACT

NYESTE, PATRICK GABOR. Training Users to Counteract Phishing. (Under the direction of Dr. Christopher B. Mayhorn and Dr. Donald H. Mershon.)

Phishing is an increasingly more prevalent form of online, social engineered scams that escalate costs and risks to society year to year. This dissertation research attempts to show an association between anti-phishing training techniques used in previous research and individual differences including: cognitive abilities (inhibition and working memory) and personality factors (Trust, Impulsivity, Computer Experience) which could affect phishing susceptibility. A thorough literature review looking at cyber-security and training includes: costs due to phishing, learning science, systems approach for training, risk, trust, online environment and cues, and recent anti-phishing training research.

ANOVA, Regression, and Signal Detection Theory (SDT) were used to ascertain individual difference and training group effects on phishing susceptibility. The individual difference research was exploratory and found working memory capacity (Alphabet Span) and Inhibition of irrelevant information (Stroop) to have an inverse relationship to phishing susceptibility. Results from a 2 (time tested) x 3 (levels of training) repeated measures design showed anti-phishing training in both a simple comic and more complex video game form is helpful in decreasing phishing susceptibility as measured by Miss rates for all individuals including college aged and computer savvy participants. Based on the results and limitations of the present study, review of the anti-phishing training literature and learning science, future training programs and future research ideas are discussed.

Training Users to Counteract Phishing

by
Patrick Gabor Nyeste

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2011

APPROVED BY:

Dr. David M. Martin

Dr. Douglas Gillan

Chair of Advisory Committee
Dr. Christopher B. Mayhorn

Co-Chair of Advisory Committee
Dr. Donald H. Mershon

DEDICATION

I dedicate this Dissertation to my Mom, Dad, Uncle Miklós, and my Fiancée, Monica, whose love and support have guided and supported me in this great endeavor.

Szívből köszönöm az összes támogatást és imák. Végül, ez kész!

BIOGRAPHY

Patrick G. Nyeste was born March 16, 1979 in Birmingham, Alabama. He attended John Carroll Catholic High School in Alabama and graduated in 1997. He then attended the University of Alabama at Birmingham and received his Bachelor of Science in Psychology with a pre-med track and minor in Chemistry in spring 2002. He then attended the University of Alabama in Huntsville and received his hybrid Master of Arts in Experimental Psychology / Human Factors in spring 2004 under the advisement of Dr. Karen Young. He was then accepted to North Carolina State University to pursue a Doctorate of Philosophy in the Human Factors and Ergonomics program within the Psychology Department. He was co-advised by Dr. Donald H. Mershon and Dr. Christopher B. Mayhorn in conducting research on a broad range of topics that included Perception, Human Factors, Training, and Cyber-Security. He graduated with a Doctorate in Philosophy in spring 2011.

ACKNOWLEDGEMENTS

This body of work would not have been feasible without the tireless, giving strength and love of many who have been on this journey with me. First and foremost, I would like to thank God and His constant support in showering me with graces to complete this important work and a mental pilgrimage of more than a fourth of my life. Second, my family in helping and encouraging me through the trials, tribulations, and joys that occupied the many years of creating this body of work. Third, my beloved fiancée's loving encouragement that has given me much joy and peace, allowing me to put great focus and heart into this dissertation. Fourth, I am indebted to Dr. Christopher B. Mayhorn's support, understanding, passion, drive, and aid in helping me find my path through this dissertation. Fifth, I wish to thank my co-chair, Dr. Donald H. Mershon, for his guidance of many years, through the lens and ears of visual and auditory Perception. Perception is a profound topic that I have enjoyed by working in his perception lab and teaching topics of how we actively experience the world to undergraduates. Sixth, I would like to thank the rest of my committee, Dr. David W. Martin and Dr. Douglas Gillan, whose comments and suggestions were very valuable in bringing this dissertation to its greatest potential. Seventh, I am grateful to Dr. Alfred A. Bartolucci for his great statistics help. Eighth, I wish to express gratitude to Kathleen DePhillips' prayers and great advice in my life before and during this dissertation. Finally, I am indebted to William Leidhe's and Brannan McDougald's tireless work in helping me with data collection and statistics help, respectively.

TABLE OF CONTENTS

List of Tables.....	vii
List of Figures.....	viii
INTRODUCTION.....	1
Personal Costs.....	2
Economic Costs.....	3
Overview of Literature Review and Systems Approach.....	5
Machine.....	9
Environment.....	9
Human.....	10
C-HIP.....	12
Situation Awareness.....	13
Risk.....	14
Machine.....	15
Environment.....	15
Human.....	15
Awareness of Risk.....	16
Trust.....	16
Attention and Cues in an online environment.....	20
Security Strategies.....	22
Pop-ups.....	25
New Types of Password Authorization.....	27
Traditional Passwords.....	27
CAPTCHAs.....	28
Picture-category based authentication.....	30
Anti-Phishing Education and Training.....	32
The Anti-Phishing computer game.....	34
Hypotheses.....	36
METHOD.....	38
Design.....	38
Participants.....	39
Apparatus.....	40
Computer Systems.....	40
Computer-Expert screening questions.....	40
Cognitive Reflection Test.....	41
Procedure.....	41
RESULTS.....	44
Attitudes and Abilities.....	44
Computer Experience.....	44

Institutional Trust.....	45
Dispositional Trust.....	45
Cognitive Reflection Test.....	45
Impulsivity.....	46
Training Information: Results from the Anti-Phishing Phil Game.....	46
Descriptive Metrics of Phishing Performance: Signal Detection Theory (SDT)..	48
Non-Parametric SDT Sensitivity and Response Bias for Phishing Performance..	50
ANOVA Analysis.....	52
Regression Analysis.....	53
DISCUSSION.....	55
Future Research.....	63
Conclusion.....	65
REFERENCES.....	67
APPENDICES.....	91
APPENDIX A. Computer-Expert Screening Questions and Internet Experience	92
APPENDIX B. Cognitive Reflection Test.....	93
APPENDIX C. Dispositional Trust.....	94
APPENDIX D. Perceived Privacy and Trust in Researchers Scale.....	95
APPENDIX E. Impulsivity (Impulse-Control).....	96
APPENDIX F. Survey Monkey example.....	97

LIST OF TABLES

Table 1	The hypotheses used in the study encompassing main effects and interactions.....	76
Table 2	Descriptive Data Table for Questionnaires and Abilities by Training Group.....	77
Table 3	Anti-Phishing Phil Game Phishing Task.....	78
Table 4	Anti-Phishing Phil Game One-Way ANOVA by Alphabet Span, Stroop, or Trust.....	79
Table 5	Signal Detection Theory Descriptives for Phishing Performance.....	80
Table 6	Non-parametric SDT sensitivity by training condition and time.....	81
Table 7	Regression Analysis by Training Group using first week Miss rate data.....	82

LIST OF FIGURES

Figure 1.	Reported dollar loss in millions due to phishing from 2004 to 2007 based on feedback of companies from the CSI Computer Crime and Security Survey.....	83
Figure 2.	An example of a CAPTCHA that a user needs to decipher.....	84
Figure 3.	The screenshot of a real email used in the study. Each email showed a From:, Subject:, and Date: as the header. At the bottom of each email the directions in bold were given to the participant.....	85
Figure 4.	The screenshot of a fake or phish email used in the study. Each email showed a From:, Subject:, and Date: as the header. At the bottom of each email the directions in bold were given to the participant.....	86
Figure 5.	Mean Values for Hit by time and training group.....	87
Figure 6.	Mean Values for Miss by time and training group.....	88
Figure 7.	Mean Values for False Alarm by time and training group.....	89
Figure 8.	Mean Values for Correct Rejection by time and training group.....	90

Training Users to Counteract Phishing

A shiny new car is being auctioned on a website you frequent multiple times a day. Over time, you have gotten used to the images and colors on the website and the general layout and design. Many times a day, you receive detailed emails from the website that show the current status of your bid for the new car. You quickly click the enclosed link in the email to go to the site so that you can make a counter-bid. The auction will end in only 5 minutes. Immediately, the auctioning site displays a page asking for your username and password including your address and billing information. Everything gets typed in quickly so you can get to the auction in time to place a bid before time runs out. You click send and immediately you are redirected to the front page of the auction site. What just happened? The website should be displaying the auction with a logged-in profile name at the top of the page. You quickly go back to the email with the embedded link and you realize that the URL address linked to a site that was misspelled and is thus located on another server. You quickly call up your credit card company and have the credit card

cancelled while cleaning out your browser cache, history, and temporary files.

Sound familiar? Unfortunately, many people have experienced a situation similar to this one. This is just one example of a rampant and socially engineered tactic called “phishing”. As a result of this type of criminal behavior, many computer users are becoming intimately familiar with the costs associated with computer security. As the following sections illustrate, these costs take a variety of forms and can range from inconsequential to catastrophic.

Personal Costs

Loss of time and increased stress levels are the immediate personal costs (Hardee, West, & Mayhorn, 2006). Long term personal costs such as decreased trust and usage of the internet for banking, shopping, and other conveniences that require personal information are also likely (Dhamija, & Tygar, 2005). Costs from the giving of personal information can have large and significant effects in the future, negatively impacting bank accounts and your identity. Canceling the credit card is a very important first step in decreasing the chance that the criminals will steal your money and identity. If this step were not undertaken in a timely fashion, money could have been lost and your credit decimated. This would have been a very large personal cost from which it might take years to recover. However, even larger economic costs have been found in our financial system and it is continuing to rise dramatically.

Economic costs

In 2005, the total lost by 313 companies to computer security attacks was \$52,494,290 with financial fraud comprising \$2,556,900, and costs due to phishing equaling \$647,510 (CSI Computer Crime and Security Survey, 2006). In 2006, the total lost by 194 companies to computer security attacks was \$66,930,950 with financial fraud comprising \$21,000, and costs due to phishing equaling \$2,757,000 (CSI Computer Crime and Security Survey, 2007). How can these monetary losses be reduced and why did the numbers increase by a factor of five or even 10 from 2005 to 2006 even with a decreasing number of respondents? In 2007, the CSI report (CSI Computer Crime and Security Survey, 2008) did not explicitly mention total loss but average dollar amount lost by a company. From these data, the possible number of phishing incidents was extrapolated based on the percentage of total loss from 2006 as shown in Figure 1.

In terms of the reduced dollar amount loss due to phishing in 2007, the number of companies reporting their monetary losses due to cyber crime has been diminishing over the last several years (shown in Figure 1 as n of respondents). However, it is very likely that phishing has increased even more based on the 325.8 percent increase in lost dollars between year 2005 and 2006. If that trend were to have continued into 2008, the amount of money lost due to phishing could have been close to nine million dollars.

On the other hand, the nine million dollar estimate could be grossly underestimating the problem. A 2007 survey by the Gartner group (Gartner, 2007) mentioned that phishing attacks caused a loss of 3.2 billion dollars, based on a sample of 4500 adults with an average

of \$866 lost per phishing occurrence. The group also found that eBay and PayPal were the most faked websites and thirty percent of malware was distributed by advertising networks. Malware is a description for software designed only for nefarious intent. Credit cards have strong fraud detection and are not as targeted as debit cards and bank accounts with weaker fraud detection. Debit cards and bank accounts have the highest rate of attacks because of the reduced fraud detection. These figures from the Gartner group show a continued increase in phishing scams and smarter, organized groups creating these scams.

A more recent article estimates that nearly one trillion dollars could have been lost due to cyber crime globally, if one includes losses of intellectual property and related damages (Mills, 2009). This research was done by antivirus and intrusion prevention software creator McAfee by assessing responses from 800 chief information officers (CIOs) from the US, United Kingdom, Germany, Japan, China, India, Brazil, and Dubai. Lost data cost the companies 4.6 billion dollars, with \$600 million spent to clean up afterwards (Mills, 2009). There appear to be varied economic losses reported from different sources that explain the vast differences in numbers. Nevertheless, the importance and need for a defense against these cyber-security and phishing attacks are stressed throughout the reports.

What follows is a discussion of some of the latest research and results on online security and phishing which Dhamija and Tygar (2005) consider a war with “battles in the user interface space.” This discussion will be grounded in a systems framework along with aspects of risk communication, trust, passwords, security strategies, and training to build up the reader’s knowledge on how to begin counteracting online security problems related to

phishing. After that discussion, the present paper will describe the proposed aspects of training users against phishing and evaluating the results based on previous research. In this effort, rather than relying on software to recognize something “phishy,” the researcher’s goal is to understand the individual differences that affect phishing susceptibility and train users to identify phishing strategies on their own. Thus, the present study’s work is consistent with an old proverb, “Give a man a fish and you feed him for a day. Teach a man to fish and you feed him for a lifetime.” Thus, give users a warning for a phishing website and their identity might survive a few more seconds. Teach users anti-phishing strategies and they will guard their identity for a lifetime.

Overview of Literature Review and Systems Approach

A big question to answer in a review of the present literature in cyber-security and social engineering schemes is what has been done in the past to counter the attacks. What are the human limitations with social and behavioral vulnerabilities and how do we guard against and reduce these risks in an online, global marketplace? The ultimate question then becomes how do users make better security decisions based on the environment influencing their expected actions (Hardee, West, & Mayhorn, 2006)? To answer these questions, the reader needs to look first at the different facets of the whole system. These can be divided into three categories: machine, environment, and human. From these elements, a computer-based training (CBT) program can be created to train users to make better security decisions. The training program should be created by analyzing a step-by-step approach within a systems framework. A later discussion of systems framework models will answer questions on where

the errors lie and on situational awareness. But first, the basis for creating a training program needs to be considered.

Mayhorn, Stronge, McLaughlin, and Rogers (2003) recommended the use of a sequential framework starting with needs assessment, person- and task-analysis, training program design, evaluation, and then provides recommendations. Needs assessment gathers information about user's goals, skills, and training. Kumaraguru et al. (2007b) mention learning science as an important factor in aiding a user's training efficacy, retention, and, most importantly, in maintaining a user's interest and attention. These learning guidelines promoted by Kumaraguru have five important elements: learning-by-doing, immediate feedback, contiguity, personalization, and a story-based agent environment. In brief, these five aspects try to ensure that a user is: getting hands-on material to learn from, receiving feedback immediately when problems arise, given pictures and words for relevant computerized instruction that happen simultaneously, receiving wording that is in a conversational style, and being guided by a character (person, comic-book hero, etc.) within a story framework. Kumaraguru's five guidelines are consistent with previous psychological and learning science research.

Learning-by-doing has a possible basis in generative learning described by Wittrock (1992). Wittrock describes the learning process as synthesizing multisensory information that includes previous knowledge to generate models of our experiences or "active generation of meaning." Therefore, learning-by-doing allows more forms of sensory information to be

utilized for a closer fit to how the brain processes learning information than by passive memorization of lists and less meaningful knowledge.

Immediate feedback is discussed by Kulhavy (1977) and Schooler & Anderson (1990) as having a positive effect on learning. The greatest positive effect on learning is a user getting feedback after an incorrect response. The present study utilized feedback immediately after a phishing email regardless of the response. This allowed the previous email to be maintained in working memory where information is simultaneously stored and processed while helping the participant understand why the email was fake. As Kulhavy mentioned, providing feedback often will promote better learning. However, Schooler & Anderson posit that participants might start depending on the immediate feedback possibly impairing learning and competing with working memory resources. These caveats will be described in more length in the discussion section.

The contiguity effect, as described by Mayer (2003), is conducive to learning by providing a close proximity to the pictures and words that are used in computer based instruction. This close proximity of pictures, words, and animations allow a greater use of working memory integrating visual graphics and verbal text models.

Personalization is explained by Mayer (2003) as having a conversation style instead of a formal third-person style for spoken and written instruction. This conversational style has been shown to engage the learner more to help integrate previous knowledge into working memory that is simultaneously organizing the visual and verbal instruction.

A story-based agent environment has its basis from peer interaction in a learning environment. Kim & Baylor (2006) describe pedagogical agents as helping to establish a social-interaction for the learner enhancing both motivation and learning in computer based instructional training. Combining these guidelines along with task and person analysis for training program design should enhance training program efficacy.

Task analysis is used to describe the requirements needed to operate a machine correctly or to perform a particular task. Person analysis shows the capabilities and limitations that the users have in the system. This analysis is even more important when the users being trained have different levels of initial experience. Training program design is based on using all the previous factors and computer based learning guidelines (Kumaraguru et al., 2007b; Sheng et al., 2007) to create the most effective program for knowledge learning and retention. This form of training bridges the gap between the successful completion of the program or task (task analysis) and the limitations and abilities of the user (person analysis). The completion of the training program leads to evaluation of the program itself to see if knowledge was retained and applied by the users at a level acceptable by the researchers. If the evaluation does not show improvement in task performance, the system framework should be revisited in an iterative fashion beginning with needs analysis until an acceptable level is reached. Thus, the discussion of the framework will start with the machines that the user will be trained to operate in the course of meeting security goals and potentially reduce phishing.

Machine

Machines have features that may change over time or may remain constant. Interacting with a computer is an ongoing process of updates and change with different forms of input and output. Web browsers change and get updated while the mouse and keyboard have remained stable for decades as input devices. Nevertheless, web browsers are being made to work on laptops, game consoles, cars, cell phones and even refrigerators with various forms of input and output types and sizes. We get feedback from the system to inform us of whether we performed our task successfully. “Good” feedback is defined as when the user gets access to the system’s status and is then able to discern the status correctly and what should be done next (Sanders & McCormick, 1993). With poor design of the system, feedback can be delayed or even absent and the user is not able to ascertain success or failure from the machine.

Environment

What cues are there in the environment to help augment or reduce demands on cognitive resources such as working memory, attention span, and speed of processing? Where is online shopping, work, and enjoyment done? What stressors are there from the boss, family, and noisy cafes? These questions start to embody the system-wide approach to a human-computer interaction paradigm. A calm environment with a spouse might be the most beneficial in bolstering working memory, attention, and speed of processing on a task. On the other hand, a noisy café that has constant distracters could lead to reduced working memory, attention span, and speed of processing ability increasing the chance of errors. Due

to individual differences and circumstances, the opposite could be true as well, with a noisy café increasing productivity. A laptop is able to be moved to various locations and can have various users and inputs. The human needs to be able to adapt to the changes in the environment and the machine. Training allows adaptation of the human to the changes of the environment and machine. What factors does the human add to this systems approach?

Human

Humans are error prone and there are many factors that can increase and decrease errors in an online computing environment. Mayhorn et al. (2003) describe the use of a person analysis to ascertain the fundamental aspects of the users of a system; it is the second step of a systems analysis for creating a computer training program for older adults. Cognitive variables such as working memory, attention span, and speed of processing, play a crucial role in our ability to process information correctly and efficiently (Czaja et al., 2006). Baddeley (1998) proposed his model of working memory that embodies attention by way of a central executive that processes and holds data through the use of two specialized processes that he refers to as slave systems. These two slave processes are the visuo-spatial sketch pad and the phonological loop. The former aspect holds visual information while the latter contains auditory and speech information. Working memory requires input from our senses for information to be processed from the environment. Even at the best of times we have difficulty with information and our environment can help or hinder information processing. Much of this difficulty with information processing is due to limited attentional resources (Wickens & Kramer, 1985); we try to make our decisions quick and efficient based on the

recognition of patterns, instead of using all our resources. This method (called satisficing by Simon, 1956) entails choosing not the optimum, but a good-enough, decision option. Nevertheless, there is plenty we do not know about the interactions between humans and the environment.

Ultimately, we need to make the machine fit with the human as much as possible through better design. However, the best design may not be possible due to funding, time, continuous changes, and a plethora of other factors. Taken as a whole, these factors lead the designers into a dynamic where they have to decide what parts of the system need to be highlighted for use and made better. Should the user be trained to fit the machine and environment, should the environment be changed to fit the machine and the human, or should the machine be changed to better fit the user and environment? To answer such questions requires models to quantitatively structure the relationship of the machine, environment, and human framework to see where the errors lie and what parts can be improved. This systems framework was mentioned previously in the overview section when describing the creation of training materials. What follows is a model of information flow designed to focus on identifying the errors and problematic spaces in all the different parts of the system not just the users. It is another method to describe what gaps exist between the user and the computer so that these areas can be filled with training and better design.

Questions involving the systems framework can be answered methodologically through observations of the interactions between human, environment, and machine. An approach to this observation is through the use of models to convey the flow of information

and error identification. The first model that will be discussed is the Communication Human Information Processing (C-HIP) which identifies how users might conceptualize risk (Conzola & Wogalter, 2001).

C-HIP

The foundations of this risk communication model rest on warning message effectiveness in a work environment. The researchers mention that C-HIP is designed to separate and diagram the various channels of warning communication flow from the system to the user. This flow of information becomes very important in determining a suitable training program that would benefit the user in protecting his or her identity with the various types of warnings available. The C-HIP model can be used to elaborate how we allocate our attention to various aspects of a website such as a warning icon or message (source) of any modality (channel) and how this information is processed (receiver). According to C-HIP, information processing for the receiver of the information is divided into steps that include attention, comprehension, beliefs and attitudes, motivation, and finally compliance behavior. This model can quickly show where the breakdowns in communication lie so that an effective intervention can be made to reduce errors (Conzola & Wogalter, 2001). However, errors in communication are not the only problem. Questions about whether the user understands and perceives the communication coming from the machine and environment are important, along with the ability to determine future consequences of the user's actions.

The C-HIP model might need to be supplemented by a component that will allow the user to make predictions of future outcomes using the flow of information between the user

and the system. Endsley proposed a Situational Awareness (SA) framework in 1988 that combined human information processing models like the newer C-HIP model to show information flow plus the projection of future status.

Situation Awareness

Thus, SA is comprised of three important elements: Perception of the elements in the current situation, comprehension of current situation, and projection of future status (Endsley, 1988). While the first two elements are consistent with C-HIP, the third, projection of future status is unique and might supplement our understanding of how users make computer security decisions. Projecting future status can be partitioned into zones of interest along a time scale. Generally, this time scale is divided into immediate, intermediate, and long-term zones which can be conceptualized as the basis of risk perception of what could happen in the future. A tool that addresses interface design and human interaction within a system framework is the Situation Awareness Global Assessment Technique (SAGAT). The satellite aspects of SAGAT that connect to SA are: environment, decision making, individual differences (ability, experience, and training), workload, pre-conceptions, and performance. The projection of future status needs to be taken into consideration through good feedback measures in the system and future consequences.

Should the credibility of the source come into question, decision making might be impacted, thereby influencing the likelihood of future compromises to the user's security. This projection could be a sizable benefit for users against phishing scams and fraud if users were aware of the potential dangers of their actions through awareness training. Thus, trust

is proportional to how much risk one is willing to take (English, Nixon, & Terzis, 2002). Therefore, a user will assign a higher trust to a system if it facilitates successful task completion while there was a higher level of risk involved (English et al., 2002). What risks are there in online browsing?

Risk

Identity theft, credit card fraud, denial of service, viruses, trojans, personal property destruction and misuse are some of the risks we face while being online. Identity theft is the stealing of a person's information and then using this information for various purposes. Newman and McNally (2005) describe the various forms of identity theft to be: "exploiting weaknesses in specific technologies and information systems, financial scams, as a motive for other crimes, facilitating other crimes, avoiding arrest, repeat victimization, and organized identity theft." (Newman & McNally, 2005, pp. 3-6) Newman and McNally also describe that credit card fraud is most often tied into identity theft and requires extensive organized networks to operate successfully. Denial of service is the mass attack on internet servers by hundreds, thousands or more computer systems that effectively push the server past its operating limit in bandwidth and resources. Thus, the server can no longer serve requests for any user. This could be one of various attacks used against a computing system that maintains security to try and find a hole in the security. Viruses, worms like Conficker (Manjoo, 2009), and trojans can be transferred to a user's system in many different ways. Nefarious websites, USB keys, and other computers on the network are just a few of the

possible routes. Many of these risks can be categorized within our machine, environment, and human framework.

Machine. Examples of risks that come from the computer (machine) aspect are often a result of security design failures, phishing design successes, personalized messages, and removable storage devices. Warning messages can have faulty designs from lack of clarity, urgency, or annoyance, and email and website design cues may give users a difficult time in judging what is real or fake (West, Mayhorn, Hardee, & Mendel, 2008). Adding environmental factors increases the risk even more so.

Environment. Time pressures, diffusion of responsibility, and multi-tasking are some of the environmental factors that shift our decision-making abilities from full attention to satisficing heuristics. Time pressures will push us to make decisions quickly, diffusion of responsibility will make us think that someone else will take care of the problem, and multi-tasking will reduce the amount of attention we give to the task at hand with finite attentional resources. Couple these risks with the limitations of our human capabilities and the problems become ever larger.

Human. Satisficing, aided by pattern recognition to arrive at an easy and quick decision, can be disastrous, with increased risk, when coupled with inattention blindness (West et al., 2008). Phishing emails and scam sites prey upon the user's quick recognition of a trusted site with even large inconsistencies being ignored due to lack of attention. Also, we have a finite amount of attentional resources, working memory, and speed of processing that can be used at any given time. These finite resources can be aided if a user's mental model

of a website is compatible with the actual model of the webpage (Kumaraguru, 2007a; Kumaraguru, 2007b; Merritt & Ilgen, 2008). The term “mental model” is accepted to originate from the work of Johnson-Laird (1983) who describes deductive problem solving, language, and internal representations of the world as our mental models. To apply this aspect to the present literature review, a mental model is an abstract representation in a user’s mind of how a system operates (Sanders & McCormick, 1993). A compatible mental model that represents the system well is considered to have utility and therefore is useful to the user. When a user’s mental model of a website does not match the website’s model, the website is considered to have poor design. Speed of processing, comprehension, and attention are negatively impacted with badly designed websites and further negatively affected with environmental stressors. These negative aspects all increase the threshold of our ability to detect warning signs for risky online computing. Thus, detection of a risk hinges on awareness of the risk before any protective action can be taken.

Awareness of Risk. Downs, Holbrook, and Cranor (2006) assessed 20 participants who were inexperienced in computer security so there would be less bias in the perception of security cues. The security cues that were rated were: following directions of suspicious email, stolen credit card number, bank account compromised, and social security number compromised which were all rated higher in security risk than large influxes of spam.

Trust

As we continue to increase our reliance on online versions of interaction with others, business, daily activities, personal finance, and e-commerce, our desire to use these systems

is contingent upon our trust of the systems. The building of online trust is different from situations involving a face-to-face encounter with other humans for the same tasks (Riegelsberger, Sasse, & McCarthy, 2005). Thus, it becomes crucial to holistically design a system to increase its trustworthiness through, not only usable design, but also understanding how the user is able to build trust within a system framework. However, phishing sites take advantage of this component to use the trust framework that has been designed to help security and transform it to serve various nefarious needs.

There are two types of trust, namely, dispositional and history-based trust for human-computer interaction. Dispositional trust is the initial level of trust a person with no prior experience has for another person or a machine. History-based trust is based on the multiple encounters a person has had with a person or machine which have then given the user changing levels of trust over time (Merritt & Ilgen, 2008). Thus, dispositional trust could be considered more stable than history-based trust when trying to predict the user's level of trust. However, history-based trust might have better predictability of individual differences than dispositional trust when a user is faced with decision making in various situations, which still needs further study (Merritt & Ilgen, 2008). Dispositional trust seems to have a direct relationship with propensity to trust within a user's personality such as an extrovert who will be more likely to trust a person or machine than an introvert (Atoyan, Duquet, & Robert, 2006; Merritt & Ilgen, 2008). Also, users are more likely to assign levels of trust to a person or device based on how they feel towards it versus what they think about it (Atoyan et

al., 2006). Merritt and Ilgen also consider trust to be affected by the interaction between the machine, environment, and the human.

How much a website can be trusted will influence a user's decision making in terms of what information a user is willing to supply to the website. The user's perception of risk will be lower for a site that is trusted. Large amounts of marketing and advertising dollars are used in creating company logos and brands that capture higher levels of trust from users (Dhamija & Tygar, 2005). This type of history-based trust encourages users to rely on logos and brands to help make their choices on what to trust. This marketed trust is then taken advantage of by phishing schemes. The logos and brands that are abused are not limited to company names and products but also encompass security authentication and domain names.

Wogalter and Mayhorn (2008) tested various website quality seals, trust seals, and domain names against fake versions and they found users to have a difficult time discerning what was actual and fictitious. The users with the greater amount of time online gave the real quality seals (VeriSign) and domain names (.edu and .gov) a higher percentage of trust significantly more than other trust seals and domain names (.org, .net, and .com). Interestingly, students, compared to non-students, gave a significantly higher level of trust to the .edu and .gov domain names and various trust seals including VeriSign. These findings illustrate that specific design features of websites and their secure interface may impact whether or not users trust these sources of information. However, security frameworks that connect websites, servers, and a user's identity are often designed with the intent of building an encompassing security system governed by security rules. Should users be allowed to see

this framework or should it be seamless? Would not the visibility of the framework increase trust of the system? Not necessarily according to Balfanz, Durfee, Smetters, and Grinter (2004).

Balfanz, Durfee, Smetters, and Grinter (2004) and Besnard and Arief (2004) discuss the trade-offs between cost and benefits in terms of costs due to security rules and benefits that are gained from usability. Balfanz et al. (2004) describe five principles for a usable security framework. First, retrofitting usability after a security design is implemented is not as good as creating a usable security system from the beginning. Second, tools are not enough to have a usable secure solution. Tools, such as security protocols and SSH, only serve as pieces of an entire security system. Third, security infrastructure and design need to be thought out at a high level in easy-to-understand pieces of information instead of security design created on a detailed and technical level. This is done so that users do not have to worry about security technology as much as what they need to have secured. Fourth, keeping the user's needs in mind should be first on the list of requirements instead of designing for security that is only usable by the creators of the system. Fifth, the system should be thought out locally in the user's environment that keeps hidden the global security framework, cryptic keys, and certificates. This emphasis on local user interaction also allows more automation in the system, so the user does not have to deal with back-end services and framework. These five points that Balfanz and colleagues (2004) have created describe an ideal security system. Many security systems are not ideal, but instead embody a variety of trade-offs.

Besnard and Arief (2004) state that usability trade-offs are often made due to high trust by users of computer systems. These trade-offs include memory problems for complex passwords, software updates for anti-virus, high trust of email attachments, and easily sharing files online are more practical than setting up a secure framework. All of these trade-offs give an immediate benefit of getting work done while not using valuable mental and time resources to setup security. Not setting up the secure framework properly allows attackers to compromise multiple levels of protection systems. Now that the user's risk and trust have been defined within an online environment context, how does the attention the user gives and the cues found in online browsing modify risk perception and trust?

Attention and Cues in an online environment

Browsers and email software have been increasing security by using various software methods called pop-up blockers, spam filters, and phishing filters (Dhamija & Tygar, 2005; Downs, Holbrook, & Cranor, 2006). Dhamija and Tygar also created a task analysis from the Anti-Phishing Working Group (APWG) that catalogues phishing attacks in a Phishing Archive. From this analysis, 10 problems for users in detecting phishing attacks were found to encompass most of the susceptibility to the attacks. Much of the susceptibility seems to stem from: a need for users to protect themselves from a hostile and insecure internet, a need to fix their access to the insecure internet without delay or they will be in trouble, and a problem with internet browsers having poor usability in terms of security. Below is a listing of the 10 problems that come from the APWG, in no particular order, which will be included in the computer-based training that were developed in the current study:

1. Users can not consistently determine sender identity in email.
2. Users can not distinguish between authentic website content and false website content based on similar look and feel.
3. Users can not dependably tell the difference between correctly formatted domain names and those that are slightly misspelled or similar in theme.
4. Users can not consistently tell the difference between hyperlinks and images of hyperlinks.
5. Users can not identify with consistency the location difference between the browser's edge (browser chrome) and the webpage content.
6. Users can not tell the difference between actual security markers on the browser chrome and security markers in the webpage content.
7. Users do not fully understand the meaning of the SSL (Secure Socket's Layer) security protocol lock icon.
8. Users can not easily notice the lack of security indicators.
9. Users can not distinguish consistently pop-up and multiple window characteristics from each other especially with similar look and feel.
10. Users do not consistently understand SSL certificates.

Compounding these inconsistencies, security icons are differentially positioned in various browsers with different designs and versions. Therefore, these problem areas need to be addressed through training, so that the user will be better able to detect and comprehend the security feedback the web browser is displaying.

Downs, et al. (2006) also found multiple cues that either trick users or do not get enough attention. These results are similar to what Dhamija and Tygar (2005) found in their task analysis of the Phishing Archive from the APWG. Downs, et al. (2006) describe examples of cues and the sensitivity of each for inexperienced computer security participants. Inexperienced computer participants embody the human aspects of the system where the user is unfamiliar with the hazards and the present and future risks are not fully known. The percentage numbers that are provided for each cue (out of 20 participants) show a very large need for training and enhancing the user's mental model for security in an online environment. These include unknown or fake *from* addresses in email (95%), secure site lock icons (85%), broken images (80%), strange domain names (55%), https (35%), letter substitutions of domain names (double v for w), and errors in emails and messages (No percentage information given for the last two). Websites might not even show the proper secure icons until data are submitted from tables and forms.

What parts of the browser do we attend to most often when seeing a new page? What parts do we attend to when we know we are on a page that needs securing (when entering credit card information) and how often do we check for those cues?

Security Strategies

Results of a study by Dhamija, Tygar, and Hearst (2006), indicate that phishing websites fooled ninety percent of participants when the designs of the site closely mimicked the legitimate site. Twenty-three percent of the participants did not look or notice the security indicators, address bars, or status bars that are found on the browser. Popup

warnings displaying that the website had an insecure certificate for proper identity were ineffective. Lastly, all participants were at risk across a range of demographic variables including hours of computer use, race, education, gender, or previous experience.

These issues of failing-to-notice or ignoring phishing indicators show up in anti-phishing software as noted by Zhang, Egelman, Cranor, and Hong (2007). Zhang and colleagues found that one out of ten of the anti-phishing tools they tested was able to detect phishing URLs correctly 90% of the time with a false alarm rate of 42% for legitimate URLs. This type of tool will simply become ignored over time and there needs to be a better way to either replace or augment these devices, so they give useable advice to consumers on whether to proceed to the website or not.

What kind of social engineering strategies are used against the user? Social engineering strategies are discussed in terms of social impact by Latané (1981) where he describes positive and negative impacts other people have over our behavior and decision making. Semantic attacks are also a form of social engineering tactics which take advantage of our gullibility and trust of technology discussed earlier in the present paper. Schneier (2000) calls these attacks the “Third wave” of computer network attacks that first started with physical damaging of computers and networks. The second wave was syntactic computer software manipulation such as Denial of Service and most recently the third wave is semantic attacks that promote mind and behavioral manipulation (Fette, Sadeh, & Cranor, 2006; Schneier, 2000). In the following examples, phishers are using the negative impacts on our behavior to acquire our personal information by giving the user a false sense of security.

Dhamija et al. (2006) mention three strategies that are used including: capitalizing on a lack of knowledge, visual deception, and capitalizing on a lack of attention. Lack of knowledge is the limited understanding of how a computer application differentiates from online programs and various internet, website and e-mail terminologies and syntaxes. This knowledge constraint also makes it difficult to distinguish between legitimate and fraudulent websites and programs.

Dhamija et al. (2006) created a set of strategy types to help define and categorize participants and their judgments after looking at a webpage. Type 1 participants only looked at security indicators within website content. Type 2 participants looked at content of the website and the domain name. Type 3 participants looked at content of the website, domain name, and the presence of HTTPS in front of the domain name. Type 4 participants looked at content of website, domain name, HTTPS, and padlock icon presence. Type 5 participants looked at content of website, domain name, HTTPS, padlock icon, and website certificate. Correct judgment scores were significantly different as a function of strategy type. The type 4 strategy seemed to have the highest success rate in judging websites correctly with the least amount of variance. On the other hand, type 1 strategies showed the least success rate in judging websites correctly with also a low amount of variance.

While it is important to insure that visible and obvious indicators are provided to assist users in determining the security of an online document, it is even more important that consistent and easily visible and interpretable indicators are available to show that an online document is not secure (Dhamija et al., 2006). Pop-up windows and warnings might be one

approach whereby a web browser can display messages about the security of a site and what state it will be in when browsing to another page (Downs et al., 2006).

Pop-ups

Pop-up warnings are dismissed due to false alarms and vague identification of what the warning is about. False alarms come from pop-ups describing a page that is not secure when the user knows it is secure. These misleading pop-up messages are shown through displays of unsigned, unverified, or expired security certificates that the user has seen on their trusted sites and therefore the user quickly dismiss the message. Vague identification becomes problematic with the usage of cryptic and confusing terminology. Words and acronyms like encryption, certificates, SSL, HTTPS, and TSL are some of the terms that are cryptic especially to novice users. On the other hand, the idea of the insecure internet can cause overreaction towards pop-up windows with users avoiding websites with pop-ups (Downs et al., 2006).

Downs et al. (2006) used four different pop-up messages with 20 inexperienced internet users to ascertain responses: if they have seen the pop-up before, would they go on to the website, stop and not go to the website, or would their action depend on other factors? The messages were: leaving a secure site, insecure form, self-signed certificate, and entering secure site. Eighty percent of the participants reported having seen at least one of the four messages when they were shown to them. However, only 50% remembered there were pop-ups before being allowed to see a website. This aspect of reduced remembrance versus recognition is based on icons and pop-up messages having ineffective comprehension when

shown many times (Downs et al., 2006). The previous research asks if the pop-ups were legitimate ones that come from the browser. What about the pop-ups that masquerade as warning messages from other sources?

Pop-ups can come from the user's operating system while browsing the internet. Users see programs crashing in their operating system and resulting pop-up warning messages are shown which appear cryptic to many. Sharek, Swofford, and Wogalter (2008) tested 42 undergraduates on four visual variants of a warning pop-up from the operating system while the participants browsed a set of 20 health websites using a Flash-created Internet Explorer 7TM browser simulator. The Flash-based browser simulator allowed a controlled environment that realistically mimicked the web browser's function while allowing data to be recorded from the participants manipulating the pop-ups. One variant of the pop-ups was the legitimate application error pop-up with a close and OK button. The other three were variants that all added minimize and maximize buttons which the legitimate one did not have. Added to that, one pop-up had a black and white flashing background instead of beige and the other added the web browser status bar to the bottom of the pop-up. Also, the mouse pointer was changed from a pointer to a hand when hovering over the OK button on the illegitimate pop-ups.

The fake pop-ups' visual variants are potential phishing indicators that are using another web browser window to mimic the operating system pop-up. The user will then press one of the buttons in an attempt to quickly close the window and this could open a download with harmful malware and viruses leading to potential identity theft. Sharek et al.

(2008) found 62% of the participants clicked on the OK button, 27% clicked on the close button, and 7% moved the pop-up away. Moreover, correctly or incorrectly responding to the pop-ups led to some interesting and rather alarming results. Sixty-five percent correctly responded to the real warning pop-up (clicking the OK button) and 73% incorrectly responded to the fake pop-up by clicking on the OK button instead of the close button. A post-task questionnaire showed that forty-two percent of the respondents who clicked on the OK button did so to “get rid of it”.

Fake pop-ups are a legitimate concern and as evidenced by previous studies, a real security hazard. The hazards appear to come from various visual cues that either trick the user into trusting the site or remain below the user’s attentional radar (Downs et al., 2006) and annoy them enough to “get rid of it” (Sharek et al., 2008). The human operators are taxed by pop-ups, because they disrupt their workflow especially if many windows are open and they are doing multiple things at once. This leads the reader to the ultimate form of pop-up, the phishing website, which can have many of the tricks that fake pop-ups have. This need for security against phishing websites, especially involving personal identity, has had many variants over the years most commonly in the password realm. Many of the phishing sites make devastating use of faking real websites to get users to relinquish passwords and enter their personal information. What is being done to help counter this?

New Types of Password Authorization

Traditional Passwords. Keeping the user’s identity secure online has burdened users with an increasing number of arcane and hard-to-remember passwords. These “strong”

passwords are harder to decipher and crack using computer algorithms. In general, strong passwords require the use of capitalized and non-capitalized letters, eight or more characters in length, including symbols, digits, and punctuation (Gehringer, 2002). These passwords are designed to help websites acknowledge that the user is legitimate (Magalhães, Revett, & Santos, 2006). However, many users have a hard time remembering these complex passwords (especially if passwords are required to be changed often) and would rather use “weaker” short ones with their spouse’s or pet’s name.

Magalhães et al. (2006) tested 60 information technology professionals on password constitution, number of persons that know one password, password change frequency, and number of passwords used often. The study found only 17% use complex passwords, most users (72%) rarely change access codes, 52% recall that at least one other person knows their password, and 65% use only 1 or 2 passwords for most of their online experience. These data are consistent with recent results reported by Nyeste and Mayhorn (2009) that indicated that 64% of computer users reported the same user name/password combinations. Moreover, Magalhães et al. (2006) mention an important contradiction in using passwords; passwords have an essential requirement to be both simple (easy to remember) yet need to be complex (secure and hard to crack). How can these disparate requirements be made attainable?

CAPTCHAs. New ways of identifying authorized users through images and category recognition seem to be better methods than just using traditional passwords. Before the reader is introduced to “passimages” or “passgraphs” the Turing Test and CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) (von Ahn,

Blum, & Langford, 2004) need to be mentioned. Alan Turing (1950) created a series of tests designed to distinguish a human and a computer using verbal queries in the form of an imitation game. Both the human and the computer would try to convince a human judge that they are each human through verbal responses to the judge's questions. A reverse or inverse Turing Test (Saygin, 2000) is given by a computer to try and distinguish if the operator is a computer or a human (von Ahn et al., 2004).

CAPTCHAs follow the inverse Turing Test through visual testing of the human user instead of verbal testing. These images are those warped looking words within grids and contorted shapes that a computer algorithm has a very hard time deciphering but can be better interpreted by the human visual perceptual system. This active process can be very frustrating to the user especially when it changes multiple times due to incorrect input and this frustration becomes a greater problem for the elderly or visually impaired. After incorrect input, the user has to re-decipher the CAPTCHA and enter the new set of characters (Figure 2).

CAPTCHAs have been used singly and as additions to passwords to insure the user giving input is human as well as filtering out the automated-scripted attacks for gaining access by "bots" and "spam-bots." Denying these "bots" is crucial, because they can be controlled by organized crime and various types of miscreants to gain personal information from social networking sites so they can better phish the user. More and more of our personal data are being stored and found online which requires semantic approaches (understanding user's behavior) to fighting phish (Fette, Sadeh, & Cranor, 2006; Schneier,

2000). Besides guarding against “bots”, CAPTCHAs are also used after failed attempts at user name and password sign-ins and during password resetting. The evolution of CAPTCHAs, the created and evolving computer algorithms that break them, and the armies of humans deciphering CAPTCHAs for monetary gain go beyond the scope of this paper (Mori & Malik, 2008). Nevertheless, CAPTCHAs serve as a segue for image interpretation as part of the user’s identification to a computer system that is more semantic in nature and less interpretable by random “bots”.

Picture-category based authentication. Vidoop (Vidoop, 2009) is a recent attempt at enabling the user to pick categories such as space, automotive, and sports as their password identification. The user need only remember what categories they picked and in what order. They then type the letters on the pictures in the appropriate order within a text field. This allows the user’s identity to be a URI (Uniform Resource Identifier) proxy for websites that use OpenID authentication. A user is identified by the same method that websites are identified with their own dot com. This address is human readable and easier to remember such as <http://phishresearcher.myvidoop.com>. MyVidoop’s implementation also requires authorization of the user’s browser to allow access with a six digit numerical password. OpenID is a free, open-source identity authentication system governed by the OpenID Foundation that has and will be incorporated into various online security providers ([OpenID](#), 2009). It also utilizes normal password authentication by other OpenID providers. Vidoop is one form of OpenID implementation where the user only has to remember one sequence of pictorial categories.

In either implementation, visual authentication or centralized password systems, greatly reduces the user's need to remember many different passwords but has the potential of increasing phishing, because one password now becomes much more valuable, allowing access to a large network of sites (Adida, 2007; Willison, 2007; Brands, 2007). The potential is there for phishing the pictorial grid (which uses three to five categories) during the period a user is redirected to an OpenID authentication provider to enter an ID and pass code which could be a phishing website.

Excluding the potential phishing negatives of OpenID, this technology appears positive to both machine (third party online security) and human (reduces password certification and remembrance) solutions to the authentication problem. Reducing the number of passwords one must remember eases the workload of the human operator significantly, because the constant authentication needed for multiple websites would be handled by the user's URI proxy and the password is created each time by categorical recognition in the case of Vidoop. Also, reducing the extent of password authentication that takes place reduces the potential for phishing by the same amount. A smaller number of passwords would also allow the user to be quicker and more robust in handling authentication in various environments. Many of the various aspects of training, the system framework, risk, trust, security strategies and finally password systems have now been discussed. A deeper background on more recent applications of training and education for anti-phishing causes needs to be discussed before leading into the present study's hypotheses.

Anti-Phishing Education and Training

Previous research by Kumaraguru et al. (2007a) on training intervention for phishing, as well as its evaluations, used embedded, non-embedded, and controlled (no formal training except an email from a friend) conditions with interesting results. The learning of phishing tricks and techniques was found to be most effective using a comic book story embedded inside of an email program, right after a user falls for a phishing attack. The embedded and non-embedded training examples used the same anti-phishing comic strip story. The embedded case has a comic immediately pop-up after the user clicked a phishing link. The non-embedded case has a comic sent as an email after the user clicked a phishing link. This knowledge of phishing techniques was retained and transferred better when learned in an embedded fashion versus the other training variations. Thus, a user could potentially have better situational awareness in predicting future phishing attacks as well as higher motivation and attention to learn through embedded training techniques. This technique would possibly change users' attitudes and beliefs about how critical it is to protect them from phishing. While Kumaraguru and colleagues tested these electronic training formats, Sheng et al. (2007) mention that in contrast to the follow-up email notification, a New York state study found that training materials given out as pamphlets were less likely to aid users in countering phishing attacks. Contextual-environmental differences between a paper pamphlet and an electronic email could at least partially explain the differences in training results.

A performance metric was used by Kumaraguru et al. (2007a) in the form of a Cognitive Reflection Test (CRT) score. This score comes from the answers to three questions designed to determine how well the user is able to ‘reflect’ and discern phishing email and to not be easily tricked. Based on Kumaraguru and colleagues’ findings, users with a high CRT are better able to guard themselves cognitively against phishing attacks. This test will be useful in the proposed study based on the merit of Kumaraguru and colleagues’ findings to help explain possible individual variations in training and phishing susceptibility. Another disturbing fact the researchers found was that 90% of users who clicked on a phishing link ended up giving their personal identification to the fake website. Thus, giving the training to the user quickly after a link is clicked should help teach them to be careful and suspicious of giving away their personal information. Interestingly, the researchers found that the users took almost three times longer on average to carefully read over the embedded condition (97 seconds) versus the non-embedded condition (37 seconds). It is highly probable that the embedded condition found the user in a state where the training had full context and maintained their attention three times longer. Thus, Kumaraguru and colleagues’ (2007a) training in the form of static images in a comic-book story format showed good results against phishing attacks.

These outcomes seem much better than the automated anti-phishing software systems and toolbars that can augment the web browser (Zhang, Egelman, Cranor, & Hong, 2007). It would seem the training allows the user to gain confidence in making the ultimate decision and not just rely on anti-phishing software that needs updating (Sheng et al., 2007).

Computer software has been used to defend users against phishing with various results shown by Zhang and colleagues (2007), but what kind of results could be expected when utilizing a more interactive form of learning such as a computer game to augment a user's decision making? Sheng et al. (2007) created and tested an anti-phishing game called Anti-Phishing Phil to help engage users in anti-phishing training. They evaluated this game against embedded methods that taught users while they checked their email, and existing training materials. The game taught users about phishing websites while the embedded method taught users about phishing emails. This game was created to help with identifying phishing URLs, what parts of the web browser would give the user cues on a webpage's trustworthiness, and how to give users the tools to check if the website is genuine through top search results (Sheng et al., 2007). The next section describes the game design while also describing three principles that were used when designing the game: the reflection principle, the story-based agent environment principle, and the conceptual-procedural principle.

The Anti-Phishing computer game

The Anti-Phishing Phil game uses a cartoon fish that is controlled by the user with guidance from a wiser and bigger cartoon fish called the "Phish Guru" who wears glasses. To help put this into the reader's context, there is a hint of Disney'sTM Finding Nemo® but without encroaching on copyrights. Before each round, there is a small comic with the Phish Guru to put the game into context as a story-based agent and train the user in a more conceptual way. The user is then brought into the round where Phil the fish needs to eat and

rejects worms based on the URL that the worm carries. The URL is shown when the user swims to a worm.

The Phish guru gives the user feedback during the game as to either why the user got “hooked” on a phish URL or why the user did well finding a legitimate URL or avoiding a bad one. This aspect helps the user learn procedural information when deciphering potentially bad URLs. Users can ask for advice from the Phish Guru at anytime when hovering over a worm so that they do not feel completely lost such as “I googled etrade.com and here is what I got: www.etrade.com.”

After completing a stage, the user is shown the reasons for why a URL was missed or correct through a table of results. By doing this, the user is allowed to reflect and learn what the answers are before moving to the next round (Reflection principle). The game has four rounds with increasing levels of difficulty in spotting various types of phishing URLs. The first is IP address URLs, the second is sub domain URLs, the third is similar and deceptive domains, and the fourth combines the previous three. These descriptions cover the basics of the game, and on top of game creation, Sheng et al. (2007) tested the game against other training techniques including embedded and existing online-anti-phishing training material.

The researchers recruited only non-experienced users who answered “no” for two of the three questions: “whether they had ever changed preferences or settings in their web browser, whether they had ever created a web page, and whether they had ever helped someone fix a computer problem.” These tests were also used effectively in other phishing studies by Kumaraguru et al. (2007a) and Downs, Holbrook, and Cranor (2006) to categorize

experienced versus inexperienced computer users. The current study used this test to differentiate the more experienced users from the inexperienced users so that individual difference information can be analyzed against phishing susceptibility and training performance outcomes.

There was no correlation between user demographics and phishing susceptibility as shown in previous research (Dhamija, Tygar, & Hearst, 2006). Thus, demographics do not appear to influence susceptibility towards phishing attacks. User performance, confidence, educational value, and fun, were rated highest for the game condition over existing training material. There was at least one condition that the researchers found to be a problem after training with the game. The users only looked for cues in the URL and if everything seemed fine, the rest of the site was trusted. Overall, Sheng and colleagues (2007) found the best results in avoiding phishing websites and identifying legitimate ones came from using the anti-phishing game.

Hypotheses

After a review of previous research, there are many articles describing the problem of phishing (Dhamija and Tygar, 2005; Dhamija, Tygar, and Hearst, 2006; Downs, Holbrook, and Cranor, 2006; Fette, Sadeh, and Cranor, 2006; Kumaraguru et al., 2007a; Kumaraguru et al., 2007b; Mills, 2009; Newman and McNally, 2005; Richardson, 2007; Richardson, 2008; Schneier, 2000; Sheng et al., 2007; Sund, 2007; Willison, 2008; Wogalter and Mayhorn, 2008; Zhang, Egelman, Cranor, and Hong, 2007) and a few that have methodologically created and tested ways to train users against phishing attacks (Kumaraguru et al., 2007a;

Kumaraguru et al., 2007b; Sheng et al., 2007; Zhang, Egelman, Cranor, and Hong, 2007). Chiefly, tests were done to see how well users learn anti-phishing techniques from three forms of training. The three forms of training are: The Anti-Phishing Phil game plus embedded training, embedded training, and no training. The proposed four hypotheses are listed in Table 1 followed by an elaboration of each hypothesis. Notice that the hypotheses are mostly non-directional as the proposed study is exploratory.

H_1 = The Anti-Phishing game plus embedded training is predicted to have the highest retention rate with the most decrease in phishing susceptibility over embedded training alone, and the no training control. Putting both the anti-phishing game and the embedded training together as a single unit will hopefully counter the negatives (cognitive tunneling towards the URL for phishing cues) that Sheng et al. (2007) found when using the Anti-Phishing game alone.

H_2 = A performance difference is predicted to be significant based on when the user will be tested on phishing susceptibility. The user will be tested on phishing susceptibility right after the training and one week afterward. The second week phishing susceptibility is expected to be higher than the first week across all training groups.

H_3 = The second week phishing susceptibility is expected to be higher than the first week across all training groups. However, the embedded and game groups are expected to have less phishing susceptibility during the second week than the control group due to training retention.

H₄ = Exploratory regression analysis will examine how individual differences including working memory, the ability to inhibit irrelevant information, impulsivity, trust, and computer experience influence phishing susceptibility. Based on the literature review, it is suspected that higher scores on either the Alphabet-span (working memory performance) or Stroop (ability to inhibit irrelevant information) tests could result in reduced phishing susceptibility. Likewise, higher scores on the Cognitive Reflection Test might also reduce phishing susceptibility. The Cognitive Reflection Test includes critical thinking math problems that are designed to have easy answers that appear to make sense at first, but are ultimately incorrect. Lower scores on Impulsivity could result in reduced phishing susceptibility. Lower scores on trust could act to reduce phishing susceptibility. Finally, higher scores on Computer Experience might show less susceptibility to phishing.

Method

Design

The proposed study utilized an experimental 2 (time tested: immediate vs. delayed) x 3 (training type: control vs. embedded vs. game plus embedded) mixed factorial design. The independent variables were time of assessment and training type. The training variable was manipulated as a between-subjects variable while time tested was a within-subject variable. Individual difference measures collected included working memory (as measured by the Alphabet-span test), inhibition (as measured by the Stroop test), impulsivity, Computer-

Experience questionnaire, dispositional trust, and trust in researchers (proxy for situational trust).

Participants

Green's rule of thumb (Green, 1991) and Cohen's (1992) power analysis was used to find the needed sample size. Based on previous research's large and medium effect sizes and the increase of variables in the present study, 84 participants were needed for a power rating of .80. Eighty-four participants (mean age = 19.5 years, SD = 2.3, range = 17-36) were recruited from the North Carolina State psychology participant pool for the main study. The number of participants is also commensurate in terms of sample size with previous research that investigated phishing, password authentication systems, and training (Kumaraguru et al., 2007a; Kumaraguru et al., 2007b; Sheng et al., 2007). There were 28 participants assigned to each training group within constraints by randomizing the training group assignment. This constraint was introduced so that not all the participants were run serially for the same group to reduce confounds due to order effects. The main study utilized training and involved procedures that assessed participant sensory capability, demographics, and performance. Sixty-five percent of the sample was female and self-report data regarding years of education completed indicated that participants had attended school for a mean of 12.7 years (SD = 1.9). Age, gender, and education were homogenous across training groups as indicated by one-way ANOVA analyses with Tukey HSD post-hoc tests comparing training group to age ($p = .214$), gender ($p = .843$), or education ($p = .912$). Thus, any results regarding training

utility are not due to any demographic confounds associated with random assignment to training conditions.

Apparatus

Computer System. A desktop Dell computer with a Pentium 4 and an LCD screen display of 1280x1024 was configured with SuperLab v.4.5 to present experimental stimuli in the form of fake and real emails running on Windows XP. Other software on the same machine included a crossword puzzle game, the Anti Phishing Phil game, and the Anti-Phishing comic. The anti-phishing game and comic were obtained from Carnegie-Mellon's CUPS (CyLab Usable Privacy and Security Laboratory). The crossword puzzle game was downloaded from Novel games and used for the control and embedded only groups to take the place of the Anti Phishing Phil game such that time in the lab was constant across groups. To reduce confounds related to playing a game before the main experiment, a crossword puzzle was chosen to approximate the searching of letters and words; this is conceptually similar to the Anti-Phishing game task of searching website addresses. In either case, the phonological loop as described in Baddeley's (1998) working memory model was used in both games. Initial pilot testing verified that the SuperLab program worked, stimuli could be read, and proper wording of instructions was clear to participants during role-playing.

Computer-Expert screening questions. Three questions (Appendix A) were used to assess general computer knowledge. The user needed to provide a "no" answer to at least two of the three questions to be considered a non-expert (Sheng et al., 2007). These three questions have been used in previous studies (Downs, Holbrook, & Cranor, 2006;

Kumaraguru et al., 2007b; Sheng et al., 2007) as valid filters for computer experience. The present study did not eliminate any participants with high computer experience. These participants were included to increase the variance and external validity of individual differences in phishing performance.

Cognitive Reflection Test. This instrument includes a set of three questions (Appendix B) designed to test the user's ability to reflect and limit their first impulse-answer that comes to mind (Kumaraguru et al., 2007a). Because the range of response for this is 0-3 with one point added per correct answer, getting all answers correct would give the user a high CRT score of 3. Such a high CRT score would mean that the user is more prone to making thoughtful judgments over snap judgments, whereas lower scores are indicative of more impulsive judgments.

Procedure

Following completion of a consent form, participants were directed to a computer in the lab where they were asked to complete an online questionnaire (labeled as questionnaire A, B, C, D and shown in Appendix A through E) created using the Survey Monkey web tool. The first section was a demographic survey (Appendix F) used to collect participant's information such as self-reported education, academic major, vision, and hearing. In the second section, the Computer-expert screening questions (Sheng et al., 2007) and the iKnow Internet Experience measure (Potosky, 2007) were used to ascertain the participant's computer experience level. In the third section, the Perceived Privacy and Trust in Researchers Scale (situational trust) (Joinson, 2007) and Dispositional Trust (International

Personality Item Pool, 2010) ascertained the participants' level of trust. Lastly in the fourth section, impulsivity was measured using The Cognitive Reflection Test (CRT) (Kumaraguru et al, 2007a) and Impulsivity assessment (International Personality Item Pool, 2010). Once the online questionnaire was completed, cognitive tests were administered in the lab. Participant's working memory was assessed using the Alphabet-span task (La Pointe & Engle, 1990). Then, a participant's ability to inhibit irrelevant information was ascertained using the Stroop test (Golden, 1978; Stroop, 1935; Melara & Mounts, 1993). Finally, vision testing was administered before anti-phishing training started. Participants were tested using a hand-held Snellen eye chart to test their vision as it corresponds to monitor distance which is approximately two feet. This tested their ability to read small text that was close which was similar to the default email size text shown on a monitor. Only the game training group had the Anti-Phishing Phil game training before the main experimental task. This game led the participant through two different training rounds with a teaching and then game section. The teaching section provided URLs and indicated what features to examine for the purpose of spotting phished emails and websites. The game section put the teaching into practice and gave the participant a score. Both the game and embedded training groups had training during the main experimental task with the use of comics that warned of the previous fake email stimuli. This comic was utilized as both an instructional aid and a feedback device for both the game and embedded training groups. No matter how the participant responded to a fake phish email, the comic would display immediately after presentation for immediate feedback.

For the main experimental task, the user role-played a friend named Bob Jones by viewing his email inbox and interacting with the email stimuli found there. These stimuli were presented on a desktop computer program, SuperLab, and participants made responses via the keyboard. Each stimulus was created using Adobe Photoshop with images and examples taken from The Anti-Phishing Work Group (2010), and PhishTank (2010), as shown in Figures 3 and 4. The email inbox included randomized email messages each with a link directing them to visit some corresponding website. Initially, the researcher read instructions to the participant. Participants used the Y and N keys to provide a “yes” or “no” response, respectively when they were tasked with answering whether they trusted each email.

During the first week, following training, the experimental task required participants to interact with 30 emails during the immediate assessment of training. Once the task was completed, the participant was scheduled to come back a week later for delayed assessment of training during which they encountered 40 emails (without any refresher training for any of the groups). Of the 40 emails encountered during the second week, 30 emails had been previously encountered during the first week to provide a measure of training retention. The remaining ten new emails were used as a measure of training transfer. In the first week, the experiment lasted a total of 45 minutes on average and the second week experiment lasted a total of 20 minutes on average. Once the second week role-play was completed, the participant was debriefed and given research credit.

Results

Given the large amount of data collected in this study, a series of analyses were conducted to test the proposed hypotheses. Below, the individual differences in terms of attitudes and abilities of the participants are described. Following the section on individual differences, results from the Anti-Phishing Training game are described including significant relationships to attitudes and abilities. Next, Signal Detection Theory is briefly mentioned and then followed by a description of a non-parametric means to examine SDT sensitivity and response bias for phishing performance. Finally, inferential statistics for ANOVAs, and regression of the results are discussed.

Attitudes and Abilities

Questionnaires measured individual differences in abilities and attitudes of each participant for computer experience, institutional trust, dispositional trust, and impulsivity. These questionnaires had scales from one to five, with one being the lowest value for experience, trust, or impulsivity, respectively. Scores of five indicated the highest values of experience, trust, and impulsivity. The cognitive reflection test had three items that were scored as incorrect (value of zero) or correct (value of one). Table 2 shows the means, standard deviations, and analysis of homogeneity for the attitudes and abilities tests broken down by training groups and overall.

Computer Experience. On average, the sample was knowledgeable in terms of computing as indicated by a mean of 3.76 (SD = .67) on the computer experience survey instrument. Thus, the sample was not composed of novices, because its members knew how

to work with a web browser, create web pages, and fix computer issues. The overall Standard Deviation of .67 indicated a small amount of variance meaning that most of the participants had a relatively high level of computer experience with mean scores across all training groups exceeding the middle value delineated by a score of 3. Computer experience did not vary by training group.

Institutional Trust. On average, the sample had somewhat high institutional trust and trust of the researchers (proxy for situational trust) keeping their information private with a mean of 4.16 ($\underline{SD} = .73$). The Standard Deviation of .73 showed the highest variance of all the attitude questionnaires. Some participants were somewhat trusting of the researchers and others were neither trusting nor distrusting. Nevertheless, the range of variation was still mostly above the middle value of three. Institutional trust did not vary by training condition.

Dispositional Trust. On average, the sample had a relatively high level of dispositional trust and a somewhat high trust of people in general but not as high as institutional trust with a mean of 3.51 ($\underline{SD} = .53$). The Standard Deviation of .53 showed a tighter range of variance when compared to institutional trust. Thus, the lower mean and tighter variance compared to institutional trust showed the sample to be slightly less trusting of people in general than the NC State researchers conducting the present study. Dispositional trust did not vary by training condition.

Cognitive Reflection Test. On average, the cognitive reflection test showed a fairly high impulsive sample using math problem solving with a mean of .37 ($\underline{SD} = .38$). Thus, the number of problems correctly solved was low and the sample tended to impulsively select the

wrong answer. The Standard Deviation of .38 showed the lowest variance and range based on the test only having three questions. Results from Levene's test of homogeneity suggest that Cognitive reflection test performance did vary by training condition such that those in the Embedded training condition produced the most impulsive scores ($M = .21$), whereas those in the Game condition produced the least impulsive scores ($M = .51$).

Impulsivity. Overall scores on the impulsivity questionnaire were near the middle of the distribution with a mean of 2.41 ($SD = .57$). This mean is based on self-reported answers that showed a more moderate level of impulsivity compared to the cognitive reflection test. On the other hand, it could mean that the lower scores in the cognitive reflection test are due to math skills and not impulsivity. The Standard Deviation of .57 showed a higher range of variation than the cognitive reflection test. Given the range of .57 and a mean score of 2.41, these results suggest that participants generally viewed themselves as moderately impulsive. Impulsivity scores did not vary by training condition.

Training Information: Results from the Anti-Phishing Phil Game

Only the participants from the Game training condition completed the Anti-Phishing Phil Game; thus, these reflect only the participation of 28 participants. The Anti-Phishing Phil Game was composed of two rounds and whenever a participant had more than 2 out of 8 misses (incorrectly identifying a URL), the round was repeated until the participant had 2 or fewer misses. Thus, performance in each round can be measured in terms of the total number of misses. Half the participants in the game training group (14 people) needed to repeat rounds to finish the game. In the first round, the total number of repeats across all

participants was 8. In the harder second round (shown by the higher number of misses in Table 3), the total number of repeats across all participants was 15. Table 3 provides a listing of mean misses per round as well as a combined mean for total misses for the entire game.

To determine how training performance might be related to variables such as working memory, inhibition, and trust, game misses by round and combination were included as a dependent variable in a set of one-way ANOVA analyses that used each of the attitude and performance metrics as independent variables and misses during training as the dependent variable. Descriptive data for attitudes and abilities are found on Table 2. Descriptive data for the Anti-Phishing Phil Game Phishing Task is found on Table 3. The inferential results are shown in Table 4 and discussed more thoroughly in the ANOVA section. There was a significant effect for alphabet span in the second round, $F(17, 10) = 9.015$, $p = .001$, and the total misses, $F(17, 10) = 3.572$, $p = .023$. Thus, a participant's working memory ability affected how many misses they had in the game. There was a significant effect for Stroop in the first round, $F(17, 10) = 5.933$, $p = .009$. Therefore, a participant's ability to inhibit irrelevant information influenced how many misses they had in the first round of the game. There was also a significant effect for institutional trust in the first round, $F(12, 15) = 3.041$, $p = .022$. Thus, trust in the present study's researchers also affected how many misses they had in the first round. Finally, there was a significant effect of dispositional trust in the second round, $F(16, 11) = 3.909$, $p = .013$. Therefore, individual differences such as working memory, ability to inhibit, and varying levels of trust generally affected how many

misses participants produced throughout training during completion of the Anti-phishing Phil Game.

Descriptive Metrics of Phishing Performance: Signal Detection Theory (SDT)

The benefit of users' training was assessed by determining how susceptible they were to phishing at two different stages: during the first week of training, and one week after the training. These data were used as metrics of training acquisition, retention, and transfer (i.e., how well the users absorbed and learned the material). Following initial training, performance on the main experimental task (Bob's email list) created a baseline of Signal Detection hits, misses, false alarms, and correct rejections based on susceptibility to the phishing attacks according to Sheng et al (2007). Sheng et al. (2007) proposed that Hits should be defined as correctly identifying phishing emails as untrustworthy. Misses are defined as incorrectly identifying the phishing emails as trustworthy. False Alarms (FA) would be demonstrated by incorrectly identifying the real emails as untrustworthy. Correct Rejections (CR) would be correctly identifying the real emails as trustworthy. Giving the participants more examples of anti-phishing data to create a baseline before the training could introduce training effects. Thus, the baseline comparison to a control condition in the present study was done between-subjects.

Descriptive data representing phishing performance is shown in Table 5 illustrating Hit, Miss, False Alarm, and Correct Rejection aspects of SDT with the full sample (n = 84) by the time of data collection (1st vs. 2nd week). The first week SDT sample consists of 30 emails showing the mean, standard error, and standard deviation. The second week SDT

sample is a combination of the first week's 30 emails with 10 new emails. The second week retention SDT sample is the same 30 emails from the first week demonstrating how well the participants remembered the same emails shown again randomly from the previous week. The second week transfer SDT sample is comprised of only 10 new emails to gauge how well the participants could transfer their knowledge and experience from the first week to emails they have never seen before. Figure 5 shows the mean values for Hit categorized by training group and time. These mean values are based on the rate participants did not trust phishing emails. The game training group had the highest hits followed closely by the embedded group for second week hit and retention. The game and embedded training groups do not show a discernable difference in first week and second week transfer. The control group continued to show lower hit rates than the other training groups in all conditions.

Figure 6 shows the mean values for Miss categorized by training group and time. These mean values are based on the rate participants trusted phishing emails. In the first week, the control group had the most Misses while embedded and game training groups had missed less. A week after training showed an inverse relationship in terms of Misses between the control and trained groups. However, note that in retention (no new emails) the control group had a higher Miss rate than in the first week. The game and embedded training groups stayed at the same first week Missed rate.

Figure 7 shows the mean values for False Alarm categorized by training group and time. These mean values are based on the rate participants did not trust legitimate emails. The control group had the largest variation in false alarm showing a larger distrust in the

second week when compared to embedded and game training groups. However, the second week retention showed a decrease in false alarm rate for the control group. The game group maintained the same false alarm rate as it did in the first week. Transferred new email material showed all training groups to have less false alarm rates as they were not being trained on the new material.

Figure 8 shows the mean values for Correct Rejection categorized by training group and time. These mean values are based on the rate that participants trusted legitimate emails. The control group trusted the legitimate emails the most out of all groups. The game group appeared to discriminate slightly better than the embedded group in the first week with a decrease of trust in the second week. Both game and embedded training groups trusted the legitimate emails less than the control group.

Non-Parametric SDT Sensitivity and Response Bias for Phishing Performance

To assess sensitivity and response bias, the present study used non-parametric SDT tests due to the homogeneity of variance assumption being violated between training groups in the alphabet span and cognitive reflection tests (MacMillan & Creelman, 2005). Based on Sheng et al. (2007), Signal Detection Theory (SDT) was successfully used to measure a user's ability to discriminate signal plus noise (phishing sites) from noise (legitimate sites). There is a sensitivity factor (A') that compares A' between the noise and signal plus noise curves that relates to how well the user can discriminate between legitimate and phishing emails. If the training has increased the difference between A' for noise and signal plus noise curves, then the training has contributed to the user's ability to discern

authentic from fake websites. However, while sensitivity is important for understanding a participant's response from SDT, Misses are more important in the anti-phishing training case as it describes phishing susceptibility. Training users to be more skeptical of fake emails should be considered a positive result of the training.

Sensitivity (A-Prime in the non-parametric use case in the present study) is also used to measure a user's decision tendency towards either "no, these are phishing emails and not trustworthy" or "yes, these are legitimate emails and trustworthy." Sheng et al. (2007), in their parametric SDT study, found that the anti-phishing Phil game increased the distance (d') over existing training material and moved the decision bias to the left and closer to 0 (A more conservative and skeptical decision bias). Existing material gave the users a reason to become more conservative (not trusting as much) which moved the decision bias towards no, phishing sites (to the left) but not as much as the game condition. The d' for Sheng et al.'s (2007) existing material did increase as well but not as much as after the anti-phishing Phil game.

In the present study, A-Prime and Bias were calculated from the combined scores of each participant and separated by training condition and time (Table 6). The A-Prime values and biases for the control group were the largest of all three training groups across all time samples. The embedded and game training groups had close A-Prime and bias values until the second week transfer distinguishing both of them from the control group. This is shown inferentially in the next ANOVA discussion section.

ANOVA Analyses

The first hypothesis predicted that the levels of phishing susceptibility would vary by training type (embedded, game, and control). In theory, the embedded and game training types should lead to a decrease in getting phished over the third level of training type (control). The game training type would show a greater decrease in getting phishing over the embedded gaming type. To help understand the direction of the ANOVA's results, see Table 5 for training group and time means. The Miss mean scores can be conceptualized as a direct metric of phishing susceptibility. A 3 (training type) x 2 (time tested) factor repeated measures ANOVA was used to test the first, second, and third hypotheses. With an alpha level of .05, a significant main effect of training was found when comparing Miss scores (trusting phish emails) across training conditions, $F(2, 81) = 6.258, p = 0.003$. A Tukey HSD posthoc test showed a significant difference between embedded and control as well as game and control training groups. However, the difference between embedded and gaming was not significant. Thus, Hypothesis 1 was partially supported because the Miss rates of the training groups varied from that demonstrated by the Control group. A one-way ANOVA was conducted between the total time taken for each participant in the main experiment against training group and impulsivity. There was no significant difference in total time between the three training groups or impulsivity.

A significant main effect for phishing susceptibility difference between times tested was found within-subjects (Hypothesis 2) such that people were more likely to be phished in the 2nd week than in the first, $F(1, 81) = 6.998, p = 0.01$. The dependent variables tested in

this hypothesis were between first week Miss and second week Miss retention (30 emails). There was predicted to be an interaction (Hypothesis 3) that would show phishing susceptibility would increase but training type would affect the phishing susceptibility increase. Specifically, it was hoped that the game or embedded training would be statistically different than the control in keeping phishing susceptibility lower in the second week. However, the interaction between times tested and levels of training was not statistically significant. Thus, Hypothesis 3 was not supported. To determine whether differences in phishing performance might be related to the differential study times associated with the navigation of the various training groups through the Bob Jones scenario, total time taken to perform the Superlab task by training group was compared. No significant differences between completion time by training group were revealed; thus, any phishing differences observed are not due to the fact that the training groups received more feedback than the control group.

Regression Analyses

To determine whether hypothesis 4 was supported, a multiple regression analysis was conducted to determine how individual differences data might impact phishing susceptibility as measured by the first week Miss rate for each training group. These individual difference measures included the following test scores entered into the regression equation by order: Computer-Expert Screening survey and iKnow measure, Institutional Trust (Situational), Dispositional Trust, Cognitive Reflection Test, Impulsivity, Alphabet Span test (working memory), and Stroop test (inhibition). This was done to gauge which test scores predicted

the outcome of phishing susceptibility within each of the training conditions. Table 2 and Table 5 show the descriptive means of the individual differences and the training group and time sample, respectively. It was predicted that users with high CRT scores and impulsivity, high CES scores and iKnow measure, good performance on the Stroop tests, and above average working memory should be the least susceptible to phishing attacks.

Thus, regression test results that investigated the relationship between individual attitudes and performance metrics on phishing susceptibility are illustrated in Table 7. Few of the hypothesized general relationships between individual differences variables achieved statistical significance across all training groups. Rather, individual differences appeared to differentially impact the performance of separate training groups in the following very specific cases: Dispositional trust (within the game training group) was marginally significant in affecting phishing susceptibility scores (Misses) in the first week, $B = -1.175$, $t(27) = -1.487$, $p = .153$. Dispositional trust also had the highest semi-partial r-squared of $-.29$ and explained a significant proportion of variance in phishing susceptibility scores, $R^2 = .164$, $F(1, 27) = 5.11$, $p = .032$. Thus, 16.4 percent of the variation in phishing susceptibility scores within the Game training condition can be explained by Dispositional trust.

Stroop test performance (within the embedded training group) significantly affected phishing susceptibility scores (misses) in the first week, $B = -.141$, $t(27) = -3.092$, $p = .006$ with a semi-partial R-squared of $-.513$. Stroop performance also explained a significant proportion of variance in phishing susceptibility scores in the first week, $R^2 = .234$, $F(1, 27)$

= 7.931, $p = .009$. Thus, 23.4 percent of the variation in phishing susceptibility scores within the Embedded training condition can be explained by Stroop performance. Impulsivity (within the embedded training group) was marginally significant, $B = -1.4$, $t(27) = -2.0$, $p = .06$ with a Semi-partial $R^2 = -.349$. However, the R^2 was not significant.

Discussion

The current research contributes to the existing psychological literature on computer security and phishing in a number of ways. Investigation of training effects as they impact the computer security-related behaviors of a sample that could be described as an intelligent, college aged, computer literate group that was quite trusting in disposition illustrated a number of interesting findings. For instance, inferential results included various effects of training and individual differences. A positive effect of training was found in reducing phishing susceptibility (using Miss as the dependant variable) and increasing awareness of fake emails. The positive effect of both the embedded and game training groups continued into the second week compared to the control group when looking at the means even though the results were not statistically significant. On the other hand, the increased awareness produced a negative effect on the embedded and game training groups as demonstrated by increased distrust of all emails including those that are legitimate (a large increase in the false alarm rate). This was shown by the response bias that became more conservative from both of the training groups when compared to control. Also, the correct rejections were lower than in the control group. In terms of how individual differences impact phishing, Stroop

scores as a measure of ability to inhibit irrelevant information had a significant inverse relationship with phishing susceptibility in the embedded training group. As the ability to inhibit increased, phishing susceptibility decreased. Working memory capacity as measured by the Alphabet span task had a significant inverse relationship with phishing susceptibility (Miss rate) in the game training group. As working memory ability increased, phishing susceptibility decreased. Finally, anti-phishing training appears to be an excellent way of reducing phishing susceptibility (Miss rate) in terms of increasing skepticism towards fake emails amongst a wide variety of individual differences. However, the training might gain greater strength in those individuals with increased inhibitory and working memory ability.

There is the possibility that the immediate feedback of the training group with the use of the comic was too immediate based on Schooler & Anderson (1990). They mentioned that immediate feedback competes with working memory. This immediacy reduces speed on test problems and doubled the amount of errors. Adding a delay in the feedback could enhance error correction and self correction. On the other hand, Schooler & Anderson still claim the use of immediate feedback as important to learning. However, when the delayed feedback is taken away, performance is increased over immediate feedback removal. Thus, a delay on feedback might have affected the results of the present study in a positive direction with better retention. In future studies, controlling for feedback delay amount could possibly help interpret the present study's increase in false alarms during and after training as well as making the training better.

Taken together, the results from this study have shown the effects of anti-phishing training to be helpful in countering the increasing threat of phishing attacks by decreasing phishing susceptibility (as measured by Miss rate). Also, this research provides results that are consistent with multiple findings from previous research in the literature review that training helps mitigate the phishing threat (Kumaraguru et al., 2007a; Kumaraguru et al., 2007b; Sheng et al., 2007; Zhang, Egelman, Cranor, and Hong 2007). As illustrated in the results of previous research (Sheng et al., 2007), the efficacy of anti-phishing training can be shown using signal detection theory to help understand the range of the participants' responses. However, the present study based much of its results on the Miss rate to gauge phishing susceptibility. Also, the present study made some modifications to the methodology used in previous studies due to the addition of extra sets of questionnaires for attitudes and individual difference testing. While many of these factors failed to disprove the null hypothesis, much can still be gleaned from the data as the Miss rate decreased due to being in the embedded or game condition and the Hit rate also increased as shown in Figures 5 and 6. However, the False Alarm rate (distrusting real emails) increased much more in both the embedded and game training groups. This shift in False Alarm could serve to explain why the embedded and game training groups demonstrated decreased sensitivity as measured with A-prime. Because SDT sensitivity utilizes the relatively less meaningful Hit Rate (distrusting fake emails) and False Alarm rate, the present study utilizes results from the more contextually important Miss Rate (Phishing Susceptibility or trusting fake emails) for the dependent variable analysis.

As predicted from previous research, the embedded email training group and game group were both significantly different from the control group. However, Sheng et al. (2007) found that the gaming group was significantly different from the embedded training while the present study found the differences between gaming and embedded conditions to not be significant. A possible reason for this change is that the present study used different techniques to gauge participant's phishing responses even though the same training material was used (six picture warnings for fake email and the Anti-Phishing Phil game). A picture of an email was used that solicited a quicker yes or no response for trustworthiness versus a high fidelity email system that Sheng et al. (2007) used to increase external validity. While these results demonstrate validity for generic anti-phishing training, more detail needs to be discussed about the present study's sample and what their attitudes and performance mean for the present study's results.

First, as described earlier in the Method section, age, gender, and education were homogenous to training group based on one-way ANOVA and supporting Tukey HSD post-hoc tests comparing training group to age ($p = .214$), gender ($p = .843$), or education ($p = .912$). This eliminates some confounds that any single training group had a significantly different sample from the others. However, this homogenous college student sample could also have compromised the external validity and range of variance that could have been possible by not using technologically savvy and trusting university students as discussed below.

Table 2 shows the listing of means for the questionnaires given to NC State undergraduate psychology students. These data are important in helping to understand what effect the sample had on the significant and non-significant results in the present study. Computer experience had a mean of 3.76 ($SD = .67$) that showed a technologically competent college sampling. In other words, most of the sample had a good deal of website knowledge, usage, and creation including fixing computer issues.

Institutional trust had the highest mean rating of 4.16 ($SD = .73$) that showed the students trusted the researchers that conducted the online survey portion of the study. Because the survey was a link from an NC State psychology participant pool website, they probably thought of the survey as being valid and safe since the survey must have been reviewed and approved by a university IRB. This result also demonstrates how source credibility can be an important factor in assessing the legitimacy of emails and related websites. Here, the participants proved to be more trusting of a familiar, valid, and presumably safe entity (NC State) much like an approved and safe company or website logo (Wogalter & Mayhorn, 2008).

Dispositional trust had a mean of 3.51 ($SD = .53$) that was lower than mean institutional trust. This type of trust gauges the tendency of the participants to trust based on more general statements (e.g., I believe that people are essentially moral). Thus, even in a more general sense, these questionnaires show that students are quite trusting and find information to be more credible, separating them from the general adult population as shown in Metzger, Flanagin, & Zwarun (2003) as well as the present study. Also, Metzger et al.

(2003) found that students check their sources less frequently (though their source checks could potentially be quite thorough given their level of access to information resources such as libraries and institutional web content) than the adult population as well.

The non-parametric signal detection theory (SDT) sensitivity results also supported this trend of trust by the participants (Table 6). The control training group started with the highest bias of .23 and stayed high (.27) during the next week's sampling of phishing and legitimate emails. Another view of SDT results, showing High trust in graph form, is also inferred from Figures 5 and 8 indicating that the control group had a lower hit rate and a higher correct rejection rate than embedded and game conditions. Both the embedded and game training group had a lower and more conservative bias range that demonstrated a decrease in trust for all emails viewed (from .11 to .17). Overall, the game condition had the lowest and most conservative bias and the absolute lowest bias when viewing the ten new emails in the second week (.07).

However, the A-prime metric for sensitivity between the noise (legitimate sites) and the signal + noise curves (phishing sites) was the opposite of what was predicted to happen based on previous research (Sheng et al., 2007). Sheng et al., (2007) used a similar (parametric) SDT method and found the game condition to have a bias close to 0 that was more conservative than the other training conditions. While the present study did have less overall A-prime in the embedded and game conditions than the control, the degree of difference was much smaller than what Sheng et al. (2007) found. The largest difference

found for A-prime sensitivity was between the second week control (.28) and the second week game (.07).

Table 5 shows a break down of SDT results by time of assessment including the effects of retention from previous emails and transfer to new emails. The two sections most directly comparable were the first week results and the second week (retention) results as they both had the exact same emails. The only difference between first week and second week (retention) was the randomized order in which the emails were shown and no training in the second week. When measuring phishing susceptibility, the Miss rate was used to quickly see what phishing emails were trusted. The second week retention results showed the Miss rate increase which was predicted based on previous research (Sheng et al. 2007). As other factors might influence decision bias in terms of an increased rate of Misses during the phishing task, an individual difference variable that was investigated in this work was the impact of impulsivity.

The last two questionnaires assessed impulsivity. The Cognitive Reflection Test (CRT) had a mean of .37 (SD = .38) and Impulsivity had a mean of 2.41 (SD = .57). Although both questionnaires purport to measure impulsivity, results diverge to suggest that impulsive behavior as measured by the CRT is more indicative than self-reports of impulsive behavior as measured by the Impulsivity survey. CRT uses three math questions that have an obvious but incorrect answer and less impulsive, deeper thought-driven correct answer. Note that the .37 mean is a rather low number demonstrating most of the participants did not reach two out of three correct responses to pass the CRT. Thus, based on this test, the sample used

in the present study was more impulsive. In contrast, the impulsivity questionnaire indicated that participants viewed themselves as more balanced in terms of their perceived level of impulsivity given an overall mean score of 2.41 that indicates that they believed many of the statements were inaccurate. The lack of reliability between these scores may simply reflect the difference between self-report measures and observed impulsive behavior or perhaps the confound of math skills.

Another measure of phishing susceptibility was used during the training task for the game condition only. This measure of phishing susceptibility is taken from the rounds of game play in the Anti-Phishing Phil game completed by 28 participants. It is a one-third subset of the total sample in the present study, but it was worth looking at to see if any significant results could be found (Table 4). Also, the means are shown in Table 3 with the second round of the game proving to be harder for all the participants with a higher Miss on average from two to three (harder URLs were used and a dangerous swimming shark in the game environment was included in the game scenario). One caveat derived from these data is that the Misses also include instances when the participants thought that a real website URL was a phishing site. Nevertheless, individual differences in performance data and attitudes could show interesting and valid effects on discriminating URL cues in this anti-phishing game training environment. Alphabet Span, Stroop, and the two aspects of trust tested (Institutional and Dispositional) were found to be significant when compared to the participant's Misses in each round or in total. Thus, working memory, speed of processing,

and trust are factors in how well a participant scores in this game and potentially how well they can detect fake and real cues in website URLs.

Future Research

This study utilized a sample size that was chosen to provide medium power based on large effect sizes that were documented in previous research. Also, the addition of a variety of tools used to measure individual differences in attitudes and cognitive ability necessitated the use of a larger sample. Nevertheless, future studies could use much larger sample sizes and less homogenous participants in terms of age, trust, and a variety of other potential confounding factors to better understand the relationships between individual differences and phishing susceptibility. Given the results reported here, anti-phishing training appears to be effective in reducing phishing susceptibility and increasing a conservative posture towards email consumption. However, certain variables such as the Alphabet Span, Dispositional Trust, and Impulsivity were almost significant in a regression test that possibly could become significant with a more variable and larger sample size. Based on the results from this study, a couple of recommendation points for future anti-phishing training programs can be established:

1. Understand the user's needs for training based on attitude and performance testing.

This understanding will allow a more precise and efficient method to know what training materials to customize for a specific user. For example, a user found to have high impulsivity, high trust, low computer experience, and low ability to inhibit irrelevant information (Stroop) might need more varieties and examples to train from

than a user with the opposite attitude and skill set. Through greater knowledge of various types of phishing, impulsivity will decrease due to familiarity, skepticism will increase for phish, computer experience will increase, and pin-pointing relevant information on a webpage will increase to distinguish between phish emails and websites and legitimate ones.

2. Utilize fun video games that give the user a reason to learn the various complexities found in email and website phishing sites.
3. Use training programs that make use of the “teaching moment” to show users an email or website that was illegitimate and the reasons why the user was phished. Optimally, a computer system that will show a working email system would be the most externally valid and best for training. However, as done in the present study, even showing pictures of emails was effective in decreasing phishing susceptibility. Based on Schooler & Anderson (1990), different kinds of feedback timing might prove useful for retention of learned anti-phishing strategies.
4. Tell the user that using their identity for anything online should be done in an environment with minimal distraction that allows focusing on secure and not secure identifiers in emails and websites. Training could be done that engages the user in detecting phishing emails and sites during different environmental distractions. For example, these distractions are actually done in the Anti-Phishing Phil computer game with a dangerous shark swimming around affecting the user’s ability to focus on phishing indicators.

5. Utilize training materials that give the largest range possible of legitimate security icons and organizations. Show how these can be faked on an illegitimate email or website and point out what fake icons and organizations have been previously used in phishing.
6. Repeat the training at least a week later with more training done for those that have the least experience, skills, and phishing susceptible attitudes. Participants did worse in general a week after training. Thus, it is crucial for the training to be done more than once to enhance the likelihood of skill retention and transfer.

Conclusion

Online security scams, phishing emails, increasing cyber-security attacks, and social engineering tactics are becoming more robust, complex, organized, and a greater threat. Higher costs in time, money, identity, and trust will continue to increase unless proper education can be given to all individuals and institutions at risk. The present study has shown that anti-phishing training in both a simple comic and more complex video game form is helpful in decreasing phishing susceptibility as measured by Miss rates for all individuals including college aged and computer savvy participants. One may hope this study opens the door for more research, not just in anti-phishing training, but in the way that individual differences may interact with phishing susceptibility. Increasing the population variability and sample size as discussed earlier might truly show very interesting results that are more likely to generalize to a heterogeneous population of computer users. Ultimately, to counter-act unforeseen new social engineering tactics, other approaches could be used in conjunction

with anti-phishing training to help users avoid phishing scams. For instance, a more holistic approach to training could instruct computer users to control impulsivity, increase computer experience, and be less trusting of familiar sources such that they are more suspicious of all communications, not just strange looking emails and websites.

References

- Anti Phishing Work Group. Retrieved Oct. 2010. <http://www.antiphishing.org/>
- Adida, B. (2008). The mobile browser as a web-based platform for identity. Retrieved from <http://www.w3.org/2007/06/mobile-ajax/papers/adida.mbajax2007-authentication.pdf>
- Baddeley A. (1998). Working memory. *Comptes Rendus de l'Academie des Sciences – Series III*, 321 (2-3), pp. 167-173.
- Balfanz, D., Durfee, G., Smetters, D.K., & Grinter, R.E. (2004). In search of usable security: five lessons from the field. *Security & Privacy, IEEE*, 2(5), pp. 19-24.
- Besnard, D., & Arief, B. (2004). Computer security impaired by legitimate users. *Computers & Security*, 23, 253-264.
- Brands, S. (2007, Aug. 22). The problem(s) with OpenID. Message posted to <http://idcorner.org>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Conzola, V. C., & Wogalter, M. S. (2001). A Communication–Human information processing (C–HIP) approach to warning effectiveness in the workplace. *Journal of Risk Research*, 4, 309-322.
- Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., et al. (2006). Factors predicting the use of technology: Findings from the center for research and

education on aging and technology enhancement (CREATE). *Psychology and Aging*, 21, 333-352.

Dhamija, R., & Tygar, J. D. (2005). The battle against phishing: Dynamic security skins.

Paper presented at the *SOUPS '05: Proceedings of the 2005 Symposium on Usable Privacy and Security*, Pittsburgh, Pennsylvania. pp. 77-88.

Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why phishing works. Paper presented at the

CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montreal, Quebec, Canada. pp. 581-590.

Downs, J. S., Holbrook, M. B., & Cranor, L. F. (2006). Decision strategies and susceptibility

to phishing. Paper presented at the *Proceedings of the Second Symposium on Usable Privacy and Security*, Pittsburgh, Pennsylvania. pp. 79-90.

Endsley, M. R., Aircraft, N., & Hawthorne, C. A. (1988). Situation awareness global

assessment technique (SAGAT). Paper presented at the *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference NAECON*, Dayton, Ohio. pp. 789-795.

Fette, I., Sadeh, N., & Cranor, L. (2006). Web security requirements: A phishing perspective,

Carnegie Mellon University. Retrieved from [http://www.w3.org/2005/Security/usability-
ws/papers/13-cmu-requirements/](http://www.w3.org/2005/Security/usability-
ws/papers/13-cmu-requirements/)

- Gartner (2007, December 17). *Gartner Survey Shows Phishing Attacks Escalated in 2007; More than \$3 Billion Lost to These Attacks*. [Press release]. Retrieved from <http://www.gartner.com/it/page.jsp?id=565125>
- Gehring, E. F. (2002). Choosing passwords: Security and human factors. Paper presented at the *Proceedings of International Symposium on Technology and Society*, Raleigh, North Carolina. pp. 369-373.
- Golden, C. J. (1978). Stroop Color and Word Test. A Manual for Clinical and Experimental Uses. Wood Dale, IL: Stelting Company.
- Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, 26, 499-510.
- Hardee, J. B., West, R., & Mayhorn, C. B. (2006). To download or not to download: An examination of computer security decision making. *Interactions*, 13, 32-37. Retrieved from <http://doi.acm.org/www.lib.ncsu.edu:2048/10.1145/1125864.1125887>
- International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences. Retrieved Oct, 2010 from <http://ipip.ori.org/>
- Johnson-Laird, P. N. (1983). *Mental models : towards a cognitive science of language, inference, and consciousness*. Cambridge, MA.: Harvard University Press.

- Kim, Y., & Baylor, A. L. (2006). A social-cognitive framework for pedagogical agents as learning companions. *Educational Technology Research and Development*, 54, 569-596.
- Kulhavy, R.W. (1977). Feedback in written instruction. *Review of Educational Research*, 47, 211-232.
- Kumaraguru, P., Rhee, Y., Sheng, S., Hasan, S., Acquisti, A., Cranor, L., et al. (2007). Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. Paper presented at the *ECrime '07: Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit*, Pittsburgh, Pennsylvania. pp. 70-81.
- Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L., Hong, J., & Nunge, E. (2007). Protecting people from phishing: The design and evaluation of an embedded training email system. Paper presented at the *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, San Jose, California, USA. 905-914.
- La Pointe, L. B., & Engle, R. (1990). Simple and complex word spans as a measure of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 1118-1133.
- Latané, B. (1981). The psychology of social impact. *American Psychologist*, 36, 343-356.
- MacMillan, Neil A., & Creelman, C. Douglas (2005). *Detection theory: a user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

- Magalhães, P. S., Revett, K., & Santos, H. D. (2006). Generation of authentication strings from graphic keys. *International Journal of Computer Science and Network Security*, 6, 240-246.
- Manjoo, F.(2009). The Worm that ate the web. Retrieved April 6, 2009, from <http://www.slate.com/id/2214970/>
- Mayer, R.E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction*, 13, 125-139.
- Mayhorn, C. B., Stronge, A. J., McLaughlin, A. C., & Rogers, W. A. (2004). Older adults, computer training, and the systems approach: A formula for success. *Educational Gerontology*, 30, 185-203.
- Melara, R. D., & Mounts, J. R. W. (1993). Selective attention to Stroop dimensions: Effects of baseline discriminability, response mode, and practice. *Memory and Cognition*, 21, 627-645.
- Metzger, M. J., Flanagin, A. J., & Zwarun, L., (2003). College student Web use, perceptions of information credibility, and verification behavior. *Computers & Education*, 41, 271-290.
- Mills, E. (2009). Study: Cybercrime cost firms \$1 trillion globally. Retrieved Jan 31, 2009, from http://news.cnet.com/8301-1009_3-10152246-83.html.

- Mori, G., & Malik, J. (2003). Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA. Paper presented at the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin. pp. 134-141.
- Newman, G. R., & McNally, M. M. (2005). *Identity theft literature review*. Paper prepared for the National Institute of Justice focus group meeting, Washington, DC. Retrieved from <http://www.ncjrs.gov/pdffiles1/nij/grants/210459.pdf>
- Nyeste, P. G., & Mayhorn, C.B. (2009). Perceptions of cybersecurity: An exploratory analysis. Proceedings of the 17th World Congress of the International Ergonomics Association. Beijing, China.
- OpenID. (2009). OpenID foundation. Retrieved Jan 23, 2009, from <http://www.openid.net>
- PhishTank. (2010). Retrieved Oct. 2010, from <http://www.phishtank.com/>
- Richardson, R. (2007). CSI computer crime and security survey. *Computer Security Institute*. Retrieved from <http://i.cmpnet.com/v2.gocsi.com/pdf/CSISurvey2007.pdf>
- Richardson, R. (2008). CSI computer crime and security survey. *Computer Security Institute*. Retrieved from <http://www.cse.msstate.edu/~cse6243/readings/CSIsurvey2008.pdf>
- Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2005). The mechanics of trust: A framework for research and design. *International Journal of Human-Computer Studies*, 62, 381-422.
- Sanders, M. S., & McCormick, E. J. (1993). *Human factors in engineering and design* (7th ed.). New York: McGraw-Hill.

- Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and Machines*, 10, 463-518.
- Schneier, B. (2000). Inside risks: Semantic network attacks. *Communications of the ACM*, 43(12), p. 168.
- Schooler, L.J., & Anderson, J.R. (1990). The disruptive potential of immediate feedback. Paper presented at the *Proceedings of the Annual Conference of the Cognitive Science Society*, Hillsdale, New Jersey. pp. 702-708.
- Sharek, D., Swofford, C., & Wogalter, M. (2008). Failure to recognize fake internet popup warning messages. Paper presented at the *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting*, New York, New York.
- Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L., Hong, J., et al. (2007). Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. Paper presented at the *SOUPS '07: Proceedings of the 3rd Symposium on Usable Privacy and Security*, Pittsburgh, Pennsylvania. pp. 88-99.
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63, 129-138.
- Stroop, J.R. (1935). Studies of interferences in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643-662.

- Sund, C. (2007). Towards an international road-map for cybersecurity. *Online Information Review*, 31, 566.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59, 433-460.
- Vidoop. (2009). VidoopSecure – Strong Authentication. Retrieved Jan 23, 2009, from <http://www.vidoop.com>
- von Ahn, L., Blum, M., & Langford, J. (2004). Telling humans and computers apart automatically. *Communications of the ACM*, 47, 56-60.
- West R., Mayhorn C. B., Hardee J. B., & Mendel J. (2008). *The weakest link: A psychological perspective on why users make poor security decisions*. In M. Gupta & S. Sharman (Eds.), *Social and Human Elements of Information Security: Emerging Trends and Countermeasures* (pp. 43-60). Hershey, PA: Idea Group, Inc.
- Willison, S. (2007). Solving the OpenID phishing problem. Retrieved from <http://simonwillison.net/>
- Wittrock, M.C. (1992). Generative learning processes of the brain. *Educational Psychologist*, 27, 531-541.
- Wogalter, M. S., & Mayhorn, C. B. (2008). Trusting the internet: Cues affecting perceived credibility. *International Journal of Technology and Human Interaction*, 4, 75-93.

Zhang, Y., Egelman, S., Cranor, L., & Hong, J. (2007). Phinding phish: Evaluating anti-phishing tools. Paper presented at the *Proceedings of the 14th Annual Network and Distributed System Security Symposium*, Banff, Alberta, Canada.

Table 1. The hypotheses used in the study encompassing main effects and interactions.

H₁ There will be a significant main effect between the levels of training on the measure of phishing susceptibility (Miss Rate).

H₂ There will be a significant main effect between the levels of time on the measure of phishing susceptibility (Miss Rate).

H₃ There will be a significant two-way interaction between time tested and level of training on the measure of phishing susceptibility (Miss Rate).

H₄ Exploratory: There might be significant influences on phishing susceptibility from individual differences in cognitive abilities and attitudes. Exploratory regression analysis will examine how individual differences including Alphabet-span, Stroop, Impulsivity, Trust, and Computer Experience influence phishing susceptibility(Miss Rate).

Table 2. Descriptive Data Table for Questionnaires and Abilities by Training Group

	<u>Control</u>		<u>Embedded</u>		<u>Game</u>		<u>Overall</u>	
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>
<u>Alphabet</u> ^a	33.0	1.33	33.7	2.70	31.7	1.83	32.8	10.7
<u>Stroop</u>	52.7	1.63	49.6	1.74	50.0	1.47	50.7	8.56
<u>Computer Experience</u>	3.82	0.11	3.83	0.14	3.63	0.13	3.76	0.67
<u>Institutional Trust</u>	4.28	0.12	4.15	0.15	4.06	0.15	4.16	0.73
<u>Dispositional Trust</u>	3.51	0.09	3.61	0.09	3.41	0.12	3.51	0.53
<u>Cognitive Reflection</u> ^b	0.37	0.08	0.21	0.05	0.51	0.07	0.37	0.38
<u>Impulsive Control</u>	2.40	0.10	2.45	0.12	2.38	0.11	2.41	0.57

^a Alphabet span did not pass Levene's test of homogeneity. $F(2, 81) = 7.238, p = .001$

^b Cognitive Reflection did not pass Levene's test of homogeneity. $F(2, 81) = 5.785, p = .004$

Table 3. Anti-Phishing Phil Game Phishing Task^a

<u>Game Round</u>	<u><i>n</i></u>	<u><i>M</i>^b</u>	<u><i>SD</i></u>
Phish 1st round	28	2.07	2.054
Phish 2nd round	28	3.00	2.944
Phish total	28	5.07	3.610

^a The anti-phishing game task was used for the Game group only.

^b Mean Misses were when the participant failed to identify correct or fake URL in the game.

Table 4. Anti-phishing Phil Game^a One-Way ANOVA by Alphabet Span, Stroop, or Trust

	<u>df</u>	<u>MS</u>	<u>F</u>	<u>Sig.</u>
<u>Alphabet Span</u>				
phish 2nd round Miss	17	12.922	9.015	.001
Within Groups	10	1.433		
phish total Miss	17	17.771	3.572	.023
Within Groups	10	4.975		
<u>Stroop</u>				
phish 1st round Miss	17	5.933	4.564	.009
Within Groups	10	1.3		
<u>Institutional Trust</u>				
phish 1st round Miss	12	6.724	3.041	.022
Within Groups	15	2.211		
<u>Dispositional Trust</u>				
phish 2nd round Miss	16	12.438	3.909	.013
Within Groups	11	3.182		

^a The anti-phishing game task was used for the Game group only.

^b Misses were when the participant failed to identify correct or fake URLs in the game.

Table 5. Signal Detection Theory Descriptives for Phishing Performance

	<u>1st Week</u>			<u>2nd Week^a</u>		
	<u>M</u>	<u>SE</u>	<u>SD</u>	<u>M</u>	<u>SE</u>	<u>SD</u>
<u>Hit</u>	8.52	.278	2.54	10.4	.462	4.23
<u>Miss</u>	6.48	.278	2.54	9.55	.461	4.22
<u>False Alarm</u>	6.24	.325	2.98	8.27	.512	4.69
<u>Correct Rejection</u>	8.76	.325	2.98	11.8	.512	4.69

	<u>2nd Week (Retention)^b</u>			<u>2nd Week (Transfer)^c</u>		
	<u>M</u>	<u>SE</u>	<u>SD</u>	<u>M</u>	<u>SE</u>	<u>SD</u>
<u>Hit</u>	7.81	.348	3.19	2.65	.146	1.34
<u>Miss</u>	7.19	.348	3.19	2.33	.146	1.34
<u>False Alarm</u>	5.62	.376	3.44	2.64	.168	1.54
<u>Correct Rejection</u>	9.39	.376	3.44	2.37	.168	1.54

^a 40 emails in total were viewed the second week; ^b and ^c are split from this and analyzed separately.

^b 30 1st week emails tested retention of training for all emails shown in the first week.

^c 10 new emails tested transfer of training not shown in first week.

Table 6. Non-parametric SDT sensitivity by training condition and time

	<u>1stwk</u>	<u>2ndwk</u> ^a	<u>2ndwk (retention)</u> ^b	<u>2ndwk (transfer)</u> ^c
<u>Control</u>				
A'	0.62	0.62	0.62	0.58
Bias	0.23	0.27	0.28	0.23
<u>Embedded</u>				
A'	0.57	0.55	0.59	0.45
Bias	0.11	0.15	0.17	0.12
<u>Game</u>				
A'	0.58	0.54	0.55	0.49
Bias	0.12	0.10	0.13	0.07

^a 40 emails in total were viewed the second week; ^b and ^c are split from this and analyzed separately.

^b 30 1st week emails tested retention of training for all emails shown in the first week.

^c 10 new emails tested transfer of training not shown in first week.

Table 7. Regression Analysis by Training Group using first week Miss rate data

<u>Control</u>	<u>B</u>	<u>SE(B)</u>	<u>β</u>	<u>t</u>	<u>Sig. (p)</u>
Alphabet Span	.08	.10	.21	.89	.38
Stroop	.03	.08	.10	.43	.67
Computer Experience	-.79	1.29	-.17	-.61	.55
Institutional Trust	.18	1.0	.03	.14	.89
Dispositional Trust	.63	1.4	.10	.44	.66
Impulsivity	-.53	1.38	-.09	-.39	.71
Cognitive Reflection Test	-.30	1.6	-.05	-.19	.85
<u>Embedded</u>	<u>B</u>	<u>SE(B)</u>	<u>β</u>	<u>t</u>	<u>Sig. (p)</u>
Alphabet Span	.00	.03	.02	.08	.94
Stroop	-.14	.05	-.58	-3.1	.006**
Computer Experience	-.04	.61	-.01	-.06	.95
Institutional Trust	-.68	.56	-.324	-1.2	.24
Dispositional Trust	.02	.91	-.00	-.02	.98
Impulsivity	-1.4	.67	-.40	-2.0	.06
Cognitive Reflection Test	3.1	1.7	.38	1.8	.09
<u>Game</u>	<u>B</u>	<u>SE(B)</u>	<u>β</u>	<u>t</u>	<u>Sig. (p)</u>
Alphabet Span	.05	.05	.23	.99	.33
Stroop	-.02	.07	-.07	-.27	.79
Computer Experience	.16	.76	.05	.21	.84
Institutional Trust	-.24	.67	-.08	-.36	.72
Dispositional Trust	-1.2	.79	-.34	-1.5	.15
Impulsivity	.06	.94	.02	.06	.95
Cognitive Reflection Test	-1.0	1.3	-.18	-.79	.44

Note. Semi-partial $R^2 = -.513$ for Stroop in Embedded Training group: Semi-partial $R^2 = -.291$ for Dispositional Trust in Game Training group. Semi-partial $R^2 = -.349$ for Impulsivity in Embedded group. * $p < .05$, ** $p < .01$

Figure 1. Reported dollar loss in millions due to phishing from 2004 to 2007 based on feedback of companies from the CSI Computer Crime and Security Survey.

Figure 2. An example of a CAPTCHA that a user needs to decipher.

From: AmericanExpress@email2.americanexpress.com
Subject: Confirmation: Payment Received
Date: November 1, 2009

Is this email trustworthy enough to follow? (Y or N)

Figure 3. The screenshot of a real email used in the study. Each email showed a From:, Subject:, and Date: as the header. At the bottom of each email the directions in bold were given to the participant.

From: accountverify@americanexpress.com
Subject: American Express Account Verification
Date: November 1, 2009

Dear American Express card holder,

We need you to immediately verify the account you have with us. You will be able to gain bonus points on your card if you act quickly.

Please click on [this link](#) so that we can start the process as soon as possible.

Thank you,
American Express

Is this email trustworthy enough to follow? (Y or N)

Figure 4. The screenshot of a fake or phish email used in the study. Each email showed a From:, Subject:, and Date: as the header. At the bottom of each email the directions in bold were given to the participant.

Figure 5. 1wk Hit had 30 emails viewed the first week. 2wk Hit had 40 emails in total that were viewed the second week. 2wk Hit-1 (retention) had 30 1st week emails that tested retention of training for all emails shown in the first week. 2wk Hit-2 (transfer) had 10 new emails that tested transfer of training not shown in first week.

Figure 6. 1wk Miss had 30 emails viewed the first week. 2wk Miss had 40 emails in total that were viewed the second week. 2wk Miss-1 (retention) had 30 1st week emails that tested retention of training for all emails shown in the first week. 2wk Miss-2 (transfer) had 10 new emails that tested transfer of training not shown in first week.

Figure 7. 1wk FA had 30 emails viewed the first week. 2wk FA had 40 emails in total that were viewed the second week. 2wk FA-1 (retention) had 30 1st week emails that tested retention of training for all emails shown in the first week. 2wk FA-2 (transfer) had 10 new emails that tested transfer of training not shown in first week.

Figure 8. 1wk CR had 30 emails viewed the first week. 2wk CR had 40 emails in total that were viewed the second week. 2wk CR-1 (retention) had 30 1st week emails that tested retention of training for all emails shown in the first week. 2wk CR-2 (transfer) had 10 new emails that tested transfer of training not shown in first week.

APPENDICES

Appendix A

Computer-Expert Screening Questions and Internet Experience

Please read each statement carefully and then answer yes or no to each question.

1. Have you ever changed preferences or settings in a web browser?
2. Have you ever created a web page?
3. Have you ever helped someone fix a computer problem?

Please read each statement carefully, and then use the rating scale below to indicate the extent to which you agree or disagree with each statement.

- 1 = strongly disagree
 2 = somewhat disagree
 3 = neither agree nor disagree
 4 = somewhat agree
 5 = strongly agree

1. If a computer problem occurs while I am using the Internet, I usually know how to fix the problem
2. I know how to create a website
3. I know some good ways to avoid computer viruses
4. I am familiar with html
5. I know how to enable and disable cookies on my computer
6. I am able to download a "plug-in" when one is recommended in order to view or access something on the Internet
7. I can usually fix any problems I encounter when using the Internet
8. I help others who are learning to use the Internet
9. I download and install software updates from the Internet when necessary
10. I regularly update my virus protection software
11. I can design a nice background and/or signature for the e-mail messages I send
12. I know what a browser is
13. I have changed the settings or preferences on my computer that pertain to my Internet access

From: Potosky, D. (2007). The Internet knowledge (iKnow) measure. *Computers in Human Behavior*, 23, 2760-2777.

Appendix B

Cognitive Reflection Test

1. A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost? ____cents (5 cents)
2. If it takes five machines 5 minutes to make five widgets, how long would it take 100 machines to make 100 widgets? ____minutes (5 minutes)
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the path to cover the entire lake, how long would it take for the patch to cover half of the lake? ____days (47 days)

Appendix C

Dispositional Trust

Please read each statement carefully, and then use the rating scale below to describe how accurately each statement describes you.

- 1 = Very Inaccurate
- 2 = Moderately Inaccurate
- 3 = Neither Inaccurate nor Accurate
- 4 = Moderately Accurate
- 5 = Very Accurate

1. Trust others.
2. Believe that others have good intentions.
3. Trust what people say.
4. Believe that people are basically moral.
5. Believe in human goodness.
6. Think that all will be well.
7. Distrust people.
8. Suspect hidden motives in others.
9. Am wary of others.
10. Believe that people are essentially evil.

From: International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences (<http://ipip.ori.org/>). Internet Web Site.

Appendix D

Perceived Privacy and Trust in Researchers Scale (proxy for Situational Trust)

Using the scale below as a guide, indicate for each item how much you agree or disagree with the statement as they relate to the survey you just completed.

- 1 = strongly disagree
- 2 = somewhat disagree
- 3 = neither agree nor disagree
- 4 = somewhat agree
- 5 = strongly agree

1. The data I have provided will be kept secure and not exploited
2. The intentions of this survey are good
3. I do not doubt the honesty of this survey or its authors
4. This survey's authors are a dependable research group
5. This survey's authors have the appropriate skills and competence to conduct online surveys
6. The authors of the survey are trustworthy
7. This survey is professional

From: Joinson, A. N. (2007). Perceived privacy and trust. Unpublished manuscript.

Appendix E

Impulsivity (Impulse-Control)

Please read each statement carefully, and then use the rating scale below to describe how accurately each statement describes you.

Response Options

- 1: Very Inaccurate
 - 2: Moderately Inaccurate
 - 3: Neither Inaccurate nor Accurate
 - 4: Moderately Accurate
 - 5: Very Accurate
-
- 1. Keep my emotions under control.
 - 2. Let others finish what they are saying.
 - 3. Demand attention.
 - 4. React intensely.
 - 5. Talk even when I know I shouldn't.
 - 6. Often make a fuss.
 - 7. Shoot my mouth off.
 - 8. Am easily excited.
 - 9. Blurt out whatever comes into my mind.
 - 10. Barge in on conversations.
 - 11. Like to gossip.

From: International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences (<http://ipip.ori.org/>). Internet Web Site.

Appendix F

Survey Monkey Example

* 1. What is your NCSU Unity ID? (i.e. jrsmith@ncsu.edu)

* 2. What is your age?

Age

* 3. What is your gender?

- ☐ Male
- ☐ Female

* 4. What is your Ethnicity:

- ☐ African
- ☐ African-American
- ☐ Asian
- ☐ Caucasian
- ☐ East Indian
- ☐ Hispanic or Latino
- ☐ Native American
- ☐ Middle Eastern
- ☐ Mixed Race
- ☐ Pacific Islander
- ☐ Other (please specify)

* 5. Are you a full-time student?

- ☐ Yes
- ☐ No

6. Are you a full-time student:

If "Yes", then what is your major (or most likely major if undeclared)

7. Are you a full-time student?

If "No," what is your current occupation/profession?

* 8. Last or Highest year of school completed:

	year
Grade School to High School	<input type="text"/>
Technical/College/University	<input type="text"/>
Other (please specify)	<input type="text"/>

* 9. Was English the first language you learned?

- ☐ No
- ☐ Yes

If "No," please name your first language:

* 10. Do you wear corrective lenses?

- ☐ Yes
- ☐ No

* 11. Have you been diagnosed with a vision problem?

If Yes, explain here

* 12. Have you been diagnosed with a hearing problem?

- ☐ Yes
- ☐ No

If Yes, explain here