Using Vigilance to Quantify Human Behavior for Phishing Risk

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Engineering & Public Policy

Casey Inez Canfield

B.S., Engineering: Systems, Franklin W. Olin College of Engineering

Carnegie Mellon University
Pittsburgh, PA

August 2016

Acknowledgements

I'd like to thank my committee members, Stephen Broomell, Wändi Bruine de Bruin, Lorrie Cranor, and Baruch Fischhoff, for their guidance and feedback. Baruch – thanks for encouraging me to pursue the important questions and supporting me even when I was full of doubts. Wändi – thanks for inspiring me to come to grad school and having great advice. Stephen – thanks for being supportive and willing to hash out half-baked ideas. Lorrie – thanks for pushing me to think critically about the applications of this work.

I'd also like to thank my colleagues and coauthors for their help and advice on my thesis and non-thesis work: Alex Davis, Alain Forget, Sarah Pearman, Jeremy Thomas, Gabrielle Wong-Parodi, Tamar Krishnamurti, Kelly Klima, and the EPP staff. In addition, I'd like to thank Serge Egelman and Nicolas Christin for their feedback on the Security Behavior Observatory project.

I was supported by a National Science Foundation Graduate Research Fellowship (1121895) and the Carnegie Mellon Bertucci Fellowship. The Security Behavior Observatory was partially funded by the NSA Science of Security Lablet at Carnegie Mellon University (contract #H9823014C0140); the National Science Foundation, Grant CNS-1012763 (Nudging Users Towards Privacy); and the Hewlett Foundation, through the Center for Long-Term Cybersecurity (CLTC) at the University of California, Berkeley.

Finally, I'd like to thank my friends for making Pittsburgh such a great place to live and my family for their love. I'd like to give a special shout-out to my grandfather, Everett Ayres, who would have loved to brag about this day.

Abstract

Phishing attacks target individuals or organizations to steal information (such as credentials) or plant malware to gain broader access to IT systems. This thesis applies research on vigilance, people's ability to detect anomalies for a sustained period, to phishing risk. I (1) measure the human component of phishing susceptibility, (2) evaluate the validity of that measurement, and (3) demonstrate an approach for applying those measurements to risk analysis and evaluating behavioral interventions.

I quantify human performance using signal detection theory (SDT) for a detection task (deciding whether a message is phishing) and a behavior task (deciding what to do about a message). As applied to phishing, SDT distinguishes between users' ability to tell the difference between phishing and legitimate emails (called *sensitivity*, or d') and bias toward identifying uncertain emails as phishing or legitimate (called *response bias*, or c). I find that users do not sufficiently compensate for their limited detection ability when choosing behaviors, despite incorporating confidence in their ability and their assessment of the consequences of errors into their decisions.

I find similar results in an initial convenience (mTurk) sample and a community sample (enrolled in the Security Behavior Observatory (SBO) study). I find weak evidence for external validity of these tasks, given no relationship between performance in the experiment and negative computer security outcomes in real life (e.g. visits to malicious URLs or presence of malicious files). These results

prompt discussion of the challenges of comparing behavior in laboratory and complex real-world settings.

Lastly, I create an analytic model for evaluating anti-phishing behavioral interventions in the face of random and spear phishing attacks. Our results suggest the value of focusing on more susceptible users, particularly when defending against random attacks. This recommendation applies even when the ability to identify poor detectors is imperfect.

Overall, this thesis bridges the vigilance and computer security literature to improve measurement of phishing susceptibility and show the value of assessing behavioral interventions in terms of signal detection theory.

Table of Contents

List of Tables

List of Figures

1. Introduction

As declared in a recent presidential executive order, "the cyber threat to critical infrastructure continues to grow and represents one of the most serious national security challenges we must confront" (Exec. Order 13636, 2013). Given that cyber threats may exploit technical vulnerabilities (e.g. insufficient internal network partitions), organizational weaknesses (e.g. being understaffed), and human shortcomings (e.g. biases in judgment) – it is important to study the full socio-technical system (Apt et al., 2004; Apt et al., 2006). Although cybersecurity is addressed through people, process, and technology improvements, less progress has been made in the 'people' domain. As a result, human behavior is typically the weakest part of a cybersecurity strategy. The research addresses a growing need to integrate human judgment and decision-making in cybersecurity (Boyce, 2011; Proctor & Chen, 2015).

Phishing attacks target individuals or organizations to steal information (such as credentials) or plant malware to gain broader access to IT systems. This thesis applies research on vigilance, the study of people's ability to detect changes in stimuli over time, to phishing risk. After a literature review (Chapter 2), I develop a task to measure the human component of phishing susceptibility (Chapter 3), which I administer to a convenience sample (Chapter 3) and a community sample (Chapter 4), comparing task and real-world performance. Using human performance estimates from these studies, I develop and apply a model for assessing the benefit-cost of implementing interventions for users of varying vulnerability (Chapter 5). I conclude by examining the scientific and practical implications of this work

(Chapter 6). Except for the literature review and conclusion, the chapters are written as self-contained articles designed for separate publication. As a result, this introductory is brief.

1.1. Thesis Overview

The objective of this research is to reframe phishing detection as a vigilance task and draw parallels to the long history of research on vigilance. I propose and evaluate measuring phishing susceptibility and the effect of behavioral interventions using classical signal detection theory (SDT). As applied to phishing, SDT distinguishes between users' ability to tell the difference between phishing and legitimate emails (called *sensitivity*, or d') and bias toward identifying uncertain emails as phishing or legitimate (called *response bias*, or c).

In Chapter 3, I measure phishing susceptibility for two interrelated tasks, detection and behavior, in an online experiment. I manipulate three task variables: (1) which task comes first, detection or behavior (Experiment 1); (2) whether participants perform both tasks (Experiment 1) or just one (Experiment 2) and (3) whether participants are told, or must infer, the base rate of phishing messages.

In Chapter 4, I use a correlational study to assess the validity of the experimental measurement from Chapter 3. Using participants and data from the Security Behavior Observatory (SBO), I evaluate (1) face validity by replicating experimental tasks from Chapter 3 in a community sample, (2) construct validity by assessing the correlation with the Security Behavior Intentions Scale (SeBIS) (Egelman & Peer, 2015), and (3) predictive validity by comparing experimental

performance to adverse outcomes experienced by users' on their home systems, namely, visits to malicious websites and presence of malicious files.

In Chapter 5, I develop and deploy risk-analytic simulations to estimate the value of behavioral interventions for users with different ability levels. These models (1) identify which users are most susceptible, (2) assess the relative risk due to the most susceptible users, (3) estimate the benefit-cost of behavioral interventions targeting users of varying ability, and (4) the sensitivity to random versus spear phishing. The parameters in these models are estimated with values taken from the research literature (for the effectiveness of interventions) and from our behavioral experiments (for individual differences in performance).

1.2. Contributions of Thesis

In summary, this thesis shows the applicability of vigilance research as a framework for understanding phishing susceptibility and evaluating anti-phishing behavioral interventions, while extending that literature in this distinctive domain – which involves an intellectually demanding task, done concurrently with users' main task (unlike, say, baggage screening, where detecting deception is the primary task). From a scientific perspective, I contribute to the understanding of how task factors (as defined by the vigilance literature) influence phishing susceptibility (Chapter 3). From a practical perspective, I show that measuring phishing vulnerability in terms of signal detection theory paints a clearer picture of the interaction between user vulnerability, effectiveness of interventions, and types of threats (Chapter 5). However, efficiently measuring phishing susceptibility in terms of SDT is still a challenge worthy of future research (Chapter 4).

2. Background

This chapter summarizes research related to (1) vigilance and (2) phishing attacks.

2.1. Vigilance

First systematically studied by Norman Mackworth (1948), *vigilance* refers to the ability to remain alert in order to detect small changes or rare stimuli over time. Both psychologists, who are interested in the nature of attention, and human factors researchers, who are interested in its implications for system design, study vigilance. In this thesis, I propose framing the detection of phishing emails as a vigilance task, given the need for sustained attention over time.

Vigilance has been studied with respect to air-traffic controllers, industrial quality control, nuclear plant operators, and other monitoring/inspection tasks in military, medical, and industrial systems (Warm, Parasuraman & Matthews, 2008). Mackworth's (1948) initial studies had the practical purpose of determining the optimal watch length for airborne radar operators engaged in submarine detection. All vigilance tasks involve searching for a signal, the event of interest, amidst noise consisting of unimportant events. Mackworth developed the famous Clock test in which participants pressed a response key when an anomaly was observed. The test used a clock with one hand and no markings. The anomaly (or signal) was when the clock hand moved the distance of 2 seconds, rather than 1 second (noise). He observed a phenomenon termed *vigilance decrement*, where performance sharply decreases, begin as soon as 5 minutes into the task.

2.1.1. Measuring Vigilance

At present, most vigilance research quantifies performance in terms of signal detection theory (SDT) (Green & Swets, 1966). SDT distinguishes between users' ability to tell the difference between a signal and noise (called *sensitivity*, or d') and bias toward identifying uncertain stimuli as signals or noise (called *response bias*, or c). SDT has been used in a wide variety of contexts, including baggage screening (Wolfe et al., 2013), sexual intent (Farris et al., 2008), medical decision-making (Mohan et al., 2012), environmental risk perception (Dewitt et al., 2015), and phishing detection (Kaivanto 2014; Kumaraguru et al. 2010; Mayhorn & Nyeste, 2012; Sheng et al., 2010; Welk et al., 2015). SDT is superior to other types of assessment, such as accuracy, because it accounts for the tradeoff between hits (correctly identifying a signal as a signal) and false alarms (incorrectly identifying noise as a signal) (see Table 2-1). Maximizing accuracy need not maximize utility (Lynn & Barret, 2014; Lynn et al., 2015). For example, an individual could perceive all emails as phishing emails, which would maximize their probability of detecting phishing emails (one possible definition of accuracy). However, this would render email a useless form of communication, which would not maximize utility – unless any missed phishing attacks could bring down a system.

Table 2-1. SDT performance measures for phishing detection.

		Response	
		"Signal"	"Noise"
Stimulus	Signal	Hit	Miss
	Noise	False Alarm	Correct Rejection

SDT assumes that both signals (e.g. phish) and noise (e.g. legitimate emails) can be represented as distributions of stimuli that vary on a decision variable (e.g.

suspiciousness). Both performance parameters are defined in Figure 2-1. The further apart the distributions, the greater the sensitivity or *d'*. The response bias, *c*, reflects how biased users are toward treating a stimulus as signal or noise. It is measured by how far their decision threshold is from the intersection of the two distributions. A negative response bias ($c < 0$) reflects a tendency to call uncertain stimuli signals, whereas a positive response bias ($c > 0$) reflects the opposite. The parameters are calculated based on the observed hits (H) and false alarms (FA):

$$d' = \Phi^{-1}(H) - \Phi^{-1}(FA)$$

$$c = -\frac{\Phi^{-1}(H) + \Phi^{-1}(FA)}{2}$$

where Φ^{-1} represents the z-transformation to convert probabilities to z-scores. Here, d' is the difference between the hits and false alarms rates and c is the negative mean of the hit and false alarm rates (Macmillan & Creelman, 2004).

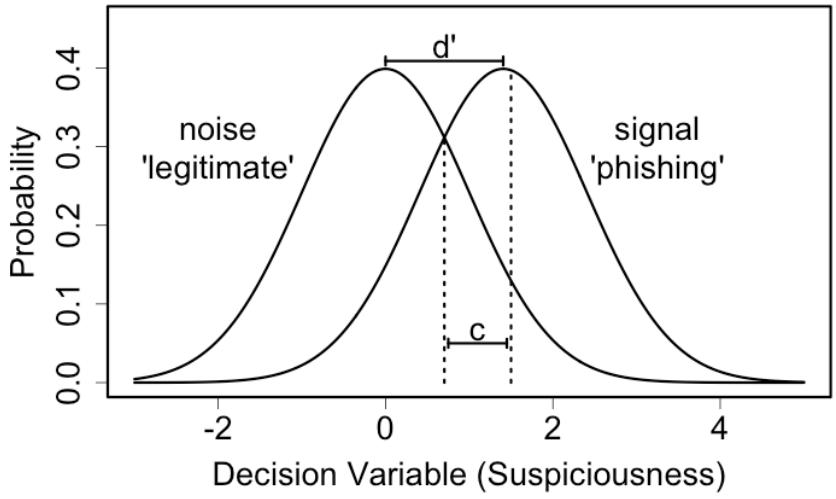

Figure 2-1. Signal detection performance parameters.

Signal detection theory (SDT) makes assumptions about the nature of perception, with the most common being (1) normality, (2) equal variance, and (3)

static d' and c. Although these assumptions are theoretically motivated, they also make the calculations more tractable. Each is discussed in more detail below.

For sensory stimuli, such as visual or auditory stimuli, there is often both theoretical and empirical justification for using Normal distributions. For example, when testing perception of auditory stimuli, tones might be random draws from a Normal distribution. In addition to the method presented above, SDT parameters can be estimated using regression, where a link function is used to identify predictors of the unobserved probability for the binary observed outcomes. Generally, SDT uses a probit link function, which gives the same results as the equations above (Knoblauch & Maloney, 2012). Several statistical tests exist to assess the appropriateness of these functions. However, all have very low power, so a null result does not necessarily indicate that the function is inappropriate (Hosmer et al., 1997). In the present research, given the lack of evidence otherwise, I assume that phishing and legitimate emails are drawn from a Normal distribution of "suspiciousness" where phishing emails are more suspicious than legitimate emails on average (see Figure 2-1). If this assumption does not hold, the estimates of d' and c may be biased, with the direction dependent on the shape of the true distribution.

Signal detection theory also assumes that perception of the noise and signal stimuli have equal variance. From a theoretical perspective, this implies that individuals are equally able to perceive both kinds of stimuli and it is the existence

of noise, rather than a feature of the signal, that leads to imperfect detection. In order to test for the equal variance assumption, we can examine the ROC curve[1]. If the ROC curve is symmetric (or slope = 1, when plotted with inverse normal coordinates), equal variance holds. An asymmetric ROC curve indicates that some assumption is violated, but not which one (DeCarlo, 1998). If the equal variance assumption is violated, the model is nonlinear (DeCarlo, 2010; Knoblauch & Maloney, 2012).

Phishing detection could be construed in two ways. One interpretation is that equal variance holds because users make mistakes in both directions because most emails contain some combination of the cues associated with phishing emails. A second interpretation is that equal variance is violated because phishing emails are specifically designed to mimic legitimate emails, so we would expect the perception of phishing emails to have higher variance. I test this assumption in the appendix for Chapter 3 (and find that equal variance holds most of the time).

Finally, SDT assumes that all noise is associated with the overlapping distributions of stimuli. This assumes that individuals have a static d' and c across trials under the same conditions. However, in reality, we might expect someone's attention to waver or be inconsistent across trials. In Chapter 3, I test for signs of

---

[1] On an ROC curve, the distance of the curve from the diagonal is described by d' while the position of a point on the curve is described by c. To draw an ROC curve, it is necessary to collect data with at least two different decision thresholds (c). This can be achieved by either (1) using a manipulation to shift c, such as changing the payoffs or base rate, or (2) using confidence ratings (Yonelinas & Parks, 2007). In most cases, it is easiest to construct an ROC curve with confidence ratings. This approach assumes that each point on the confidence scale reflects a different c. The curve formed by these points is the ROC curve. An

vigilance decrement by calculating d' and c separately for the first and second half of stimuli (and find no evidence of d' and c systematically changing over the course of the experiment).

2.1.2. Predicting Vigilance

Researchers have found that vigilance performance can be influenced by task, environmental and individual factors, as well as the interactions between them (Ballard, 1996). Figure 2-2 highlights factors that have been found to influence vigilance, and which could also be relevant for phishing detection. The factors are discussed in greater detail in the following sections.

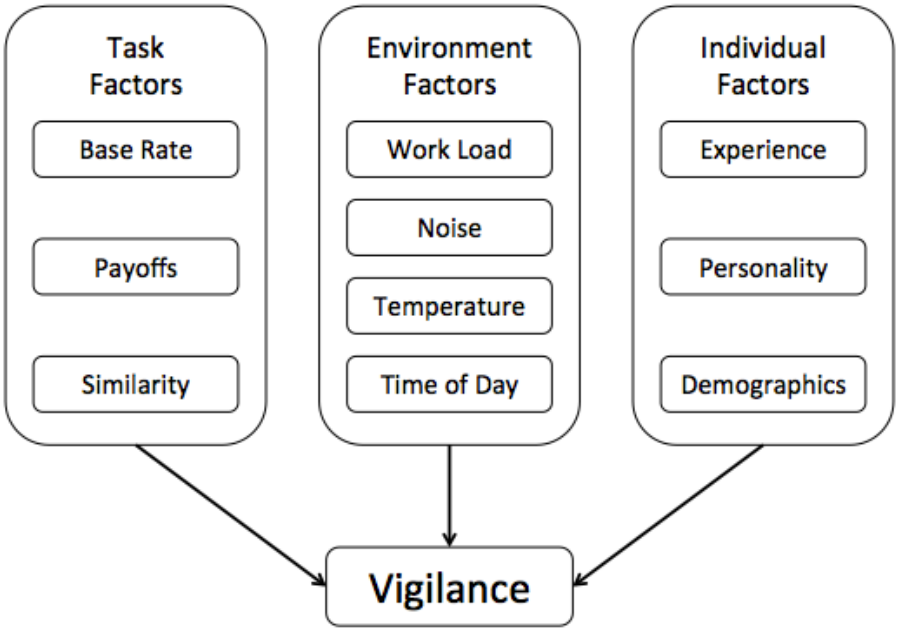

Figure 2-2. Vigilance performance is influenced by task, environment and individual factors.

2.1.2.1. Task Factors

For complex tasks, performance can be influenced by the base rate, payoffs, and similarity of stimuli (Lynn & Barrett, 2014). In general, people are more likely to

identify a stimulus as noise for low base-rate events, where it is unlikely to be a signal (Lynn & Barrett, 2014; Maddox, 2002; Navalpakkam et al., 2009; Wolfe et al., 2007). Conversely, people are more likely to identify a stimulus as a signal when missing a signal is costly and a false alarm is less costly (Lynn & Barrett, 2014; Maddox, 2002; Navalpakkam et al., 2009). In detection tasks without a clear payoff structure, participants typically try to maximize accuracy (Maddox, 2002). In addition, people are able to adapt their decision-making strategy when similarity is high (and the difference between signals and noise is very low) (Lynn & Barrett, 2014). High similarity may arise from perceptual factors (e.g. uncertainty about the difference between signals and noise), as well as environmental ones (e.g. navigating a dimly lit room)

However, not all of these factors influence performance equally. In studies, people are more sensitive to the signal base rate than to the payoffs (Maddox, 2002; Navalpakkam, Koch, & Perona, 2009). One explanation is that the base rate is typically more observable than payoffs, so people are better able to respond to it. In studies where people receive feedback, payoffs are more influential (Navalpakkam, Koch, & Perona, 2009).

An intervention called "signal injection and performance feedback" has been found to be an effective intervention for improving vigilance for sonar watchstanding (Mackie et al., 1994), baggage security screening (Wolfe et al., 2007; Wolfe et al., 2013), and medical diagnosis (Birdwell & Wolfe, 2013). In this type of intervention, signals are artificially injected throughout the task (to increase the base rate) and detectors are given feedback to indicate whether or not they

correctly detected the signal. For example, Wolfe et al. (2007; 2013) found that exposing baggage screeners to brief bursts of training at a high base rate with full feedback improved detection after they returning to the real world of a low base rate without feedback. Injecting signals artificially increases the base rate and feedback makes that increase observable, leading people to adopt a lower c and perceive more stimuli as signals (Goodie & Fantino, 1999; Kluger & DeNisi, 1996).

2.1.2.2. Environmental Factors

In general, environmental factors add stress, which interferes with performance in detection tasks. For complex cognitive tasks, white noise tends to reduce performance regardless of volume, while findings for intermittent noise (e.g. music or voices) are inconsistent. Performance is also reduced in the presence of uncomfortable ambient conditions, such as extreme hot and cold temperatures (Ballard, 2008). Some variables may have a physiological effects; for example, there is evidence that vigilance is reduced after lunch (Smith & Miles, 2007).

In some cases, the vigilance task itself is structured in ways that impose stress, reducing performance. For example, in many environments, alarms are used to attract the attention of an operator to an anomaly. This is particularly common for complex systems, where humans cannot simultaneously monitor all components. However, most alarm systems are not designed from a human factors perspective and may reduce attention without providing useful diagnostic information (Stanton, Booth, & Stammers, 1992). For example, in complex systems alarms may be correlated, so that many alarms go off at once if an event occurs (Perrow, 2011). Thus, it is possible that, for phishing detection, warnings may have

a counter-intuitive effect, which increases stress and reduces overall performance (e.g. by increasing false alarms).

2.1.2.3. Individual Factors

Individual factors found to affect performance include experience, personality and demographics. Individuals with more experience tend to have a similar d' as novices, but vary in terms of c. In air-traffic control, experienced individuals have a lower c, suggesting that they fear misses more than false alarms (Bisseret, 1981). The same effect was observed in a hazardous driving simulation (Wallis & Horswill, 2007). However, expertise had little effect in the context of detecting cyber attacks (Ben-Asher & Gonzalez, 2015). Personality has tended to have a weak relationship to vigilance performance. However, personality may be correlated to sensitivity to stress, which may reduce performance (Shaw et al., 2010). Performance may also be sensitive to affective state (i.e. feelings of happiness or sadness). For example, experiencing unpleasant affect increased attention to the base rate (Lynn et al., 2012). In terms of demographics, vigilance tends to decrease with age and there is little evidence of systematic gender effects (Ballard, 1996).

2.2. Phishing Attacks

Phishing is among the top cyberattack vectors (Symantec, 2016; Verizon, 2016). These attacks seek to trick users into thinking an email or website is legitimate, hoping to convince them to divulge sensitive information (e.g., usernames, passwords, credit card numbers) or inadvertently install malware, by clicking on malicious links or attachments. Spear phishing attacks use personal

information (e.g. known contacts, industry language, victims' names) to design more realistic and persuasive messages. Depending on the level of deception involved, it can be difficult to screen such messages automatically. As a result, human judgment plays a role in all cybersecurity systems and, by many accounts, is its weakest link (Boyce et al., 2011; Cranor, 2008; Hong, 2012; Proctor & Chen, 2015; Werlinger et al., 2009). Below, we describe (a) how phishing works, (b) phishing susceptibility, and (c) anti-phishing interventions.

2.2.1. How Phishing Works

A typical phishing attack has 4 steps (summarized in Figure 2-3). First, attackers plan and set-up the attack. Second, victims receive a malicious email. Third, victims fall for the message and take the suggested action (e.g., clicking on the link or attachment, providing their credentials). Lastly, the attackers monetize the information they received. We review these steps in greater detail below. The present research focuses on phishing emails, but attacks may occur over instant messenger, social media, VOIP, or text message (Symantec, 2016). Although the technical details of how to execute these attacks vary, the principles remain the same.

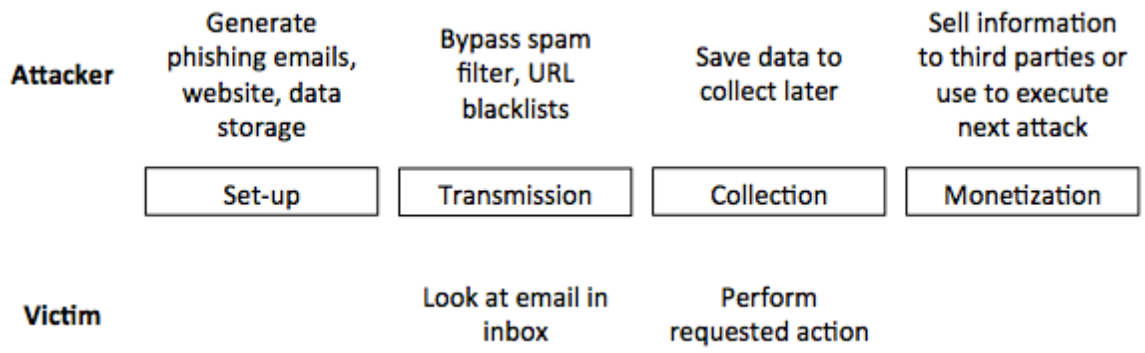

Figure 2-3. Actions required for a successful attack by both attackers and victims.

The planning and set-up of an attack vary widely. Attackers range in terms of skill and resources, from a disgruntled employee to a highly organized and well-funded nation state. Attackers must generate persuasive emails, a malicious website to steal information (or disguise malware in an attachment), and ensure that all the stolen information is properly stored to monetize later. As phishing has increased, so have efforts towards automation, which have decreased the amount of technical knowledge required. It is possible to purchase "phishing kits" that provide all of the code and materials needed to launch a large-scale phishing attack. Some of these kits are even available for free, although most free kits have backdoors in place to steal information from novice attackers (Cova, Kruegel & Vigna, 2008). In addition, attackers must gather email addresses of potential victims. Email address can be purchased in bulk from underground markets (Franklin et al., 2007), stolen directly, or collected in a simple web search.

Second, users receive phishing emails. However, not all emails that attackers send ultimately reach victims. Spam filters block emails based on their content and blacklists of known malicious URLs. For example, terms like "viagra" and "click

here" would arouse suspicion. Most spam filters examine the ratio of suspicious to total text to determine whether an email is legitimate. Attackers attempt to evade spam filters by generating new URLs and avoiding generic text (Moore & Clayton, 2007). In addition to bypassing the spam filter, attackers must ensure that their malicious website is not blocked. A group called the "rock-phish gang" evaded this issue by continuously generating new domains that redirect to each other (Moore & Clayton, 2007).

Third, users fall for the attack. The success of phishing attacks can be explained in part by dual-process theory, which posits that humans have two systems of reasoning: System 1, which makes quick intuitive judgments, and System 2, which makes conscious deliberate judgments (Kahneman, 2011; Sloman, 1996). Attackers design phishing emails, in terms of content and aesthetics, to encourage System 1, rather than System 2, processing. In the email and website, attackers invoke urgency cues in both the subject and the main text, threaten penalties for inaction, use believable senders, and use visual cues such as company logos to encourage victims to have an affective response and use System 1 (Wang et al., 2009; Wang et al., 2012; Wright et al., 2009). An example of a phishing email, using the cues discussed above, is shown in Figure 2-4.

Lastly, attackers monetize attacks. Sensitive information (e.g. credit card numbers, credentials, contact information, account information) can be bought and sold in underground markets (Franklin et al., 2007). Alternatively, the information gained in a phishing attack may be the first step in a larger orchestrated attack. However, not all attacks are driven by financial motivation. A phishing attack may

also be used to cause embarrassment or damage reputation, which may have financial consequences for the victim, but does not provide financial reward for the attacker.

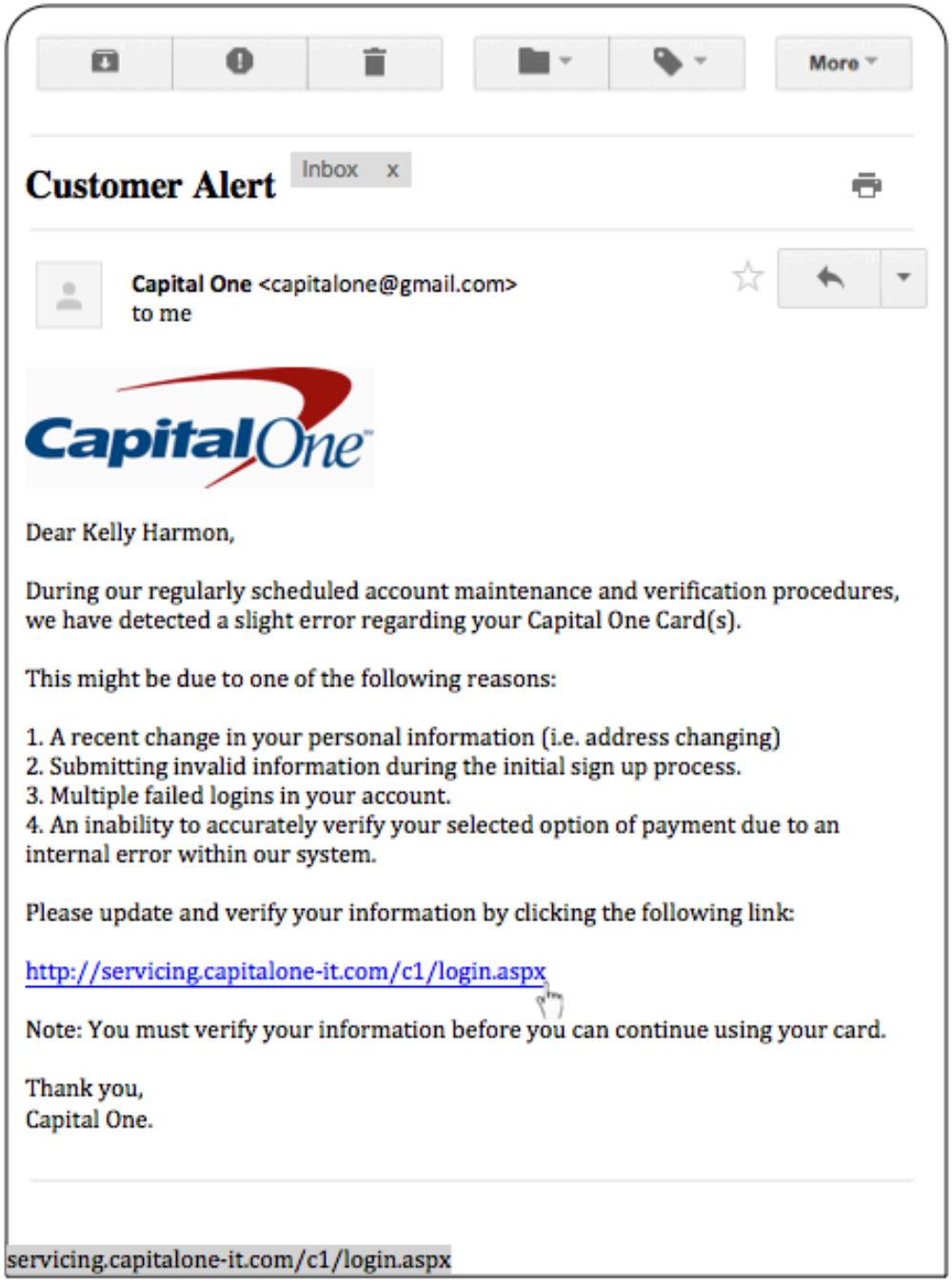

Figure 2-4. Example phishing email.

2.2.2. Phishing Susceptibility

Observational and laboratory studies suggest that the factors that influence phishing susceptibility roughly parallel the factors identified in the vigilance literature. However, little phishing susceptibility research has focused on task factors, such as base rate, payoffs and similarity. Studies have asked about phishing detection in different ways, with some directly asking whether an email or website is phishing and others asking what action would be performed. However, no studies have directly compared these types of tasks. Chapter 3 of this thesis addresses this question, calling the former "detection" and the latter "behavior." Different types of phishing attacks, for example spear phishing, may increase the similarity between phishing and legitimate emails, which reduces detection performance.

One environmental factor relevant to phishing detection is what other tasks individuals are performing. Users who receive many emails and check their email as a habit, without much conscious effort, are more vulnerable (Vishwanath et al., 2011; Vishwanath, 2015). So are users who multi-task while checking their emails or work under strict deadlines, which encourage a cursory review of emails. Platforms such as mobile devices, which offer weaker security and reduced access to cues, may also increase phishing susceptibility.

Wright & Marett (2010; Wright et al., 2009) divide individual factors into dispositional and experiential ones. In terms of dispositional factors, users who are willing to invest cognitive effort and tend to be suspicious are less susceptible (Pattinson et al., 2012; Sheng et al., 2010; Welk et al., 2015; Wright & Marett, 2010). For experiential factors, general familiarity with computers and computer security

is associated with lower susceptibility (Pattinson et al., 2012; Sheng et al., 2010; Vishwanath et al., 2011). Phishing messages use deceptive and persuasive messaging to victimize users by imposing a sense of urgency, using social cues (e.g. spoofing a known contact) and invoking authority (Butavicious et al., 2015; Dhamija et al., 2006; Luo et al., 2013). However, knowledge and experience moderate the effects of those cues (Wang et al., 2012). These factors can also interact with demographic factors; for example, women tend to have less computer knowledge and younger people tend to be less risk-averse, both of which may make them more vulnerable (Sheng et al., 2010).

2.2.3. Anti-Phishing Interventions

We can distinguish technical and behavioral interventions to reduce phishing risk. Technical interventions aim to either reduce user exposure to phishing or reduce the value of phishing attacks by making passwords less valuable (Hong, 2012; Purkait, 2012). System administrators can reduce user exposure to phishing emails by using spam filters, which divert phishing emails before users see them. Browsers employ blacklists, which block known malicious websites. However, these interventions may inadvertently increase susceptibility by artificially reducing the perceived base rate of phishing. If users see fewer malicious emails or do not realize that they clicked on a malicious website because of an uninformative error message, then they may rationally adjust their response bias (c) to account for the low base rate. Alternatively, users may believe they have nothing to fear because their machine is protecting them. In reality, blacklists and spam filters are far from perfect (Sheng et al., 2009). In addition, it is possible to make passwords less

valuable by employing two-factor authentication, which requires a physical device such as a cell phone in addition to a password, and password managers, which reduce the use of the same password across multiple websites. Evaluating the benefits of such interventions requires quantitative estimates of users' performance, the topic of this dissertation.

From a vigilance perspective, we can classify behavioral interventions as those that primarily serve to change d' or c. Interventions that increase attention or effort increase d'. Parsons et al. (2015) observed that telling users that they were being evaluated on their phishing detection ability (priming) increased their sensitivity (measured as A') without changing response bias (measured as B"). Sheng et al. (2007) developed a game called Anti-Phishing Phil to teach users how to identify phishing emails in a fun and engaging way. In pre- and post-tests, this intervention increased d', likely because users were better able to pay attention to the correct cues. However, although no change in c was observed in a laboratory setting, c increased in a field trial (Kumaraguru et al., 2010).

Signal injection and performance feedback has also been an effective intervention in the context of phishing. Kumaraguru et al. (2010) developed an intervention called "embedded training," which sends fake phishing emails to users. If they fall for fake attacks and click the link or open the attachment, they receive feedback in the form of training about phishing emails. This represents partial, rather than full, feedback since users do not receive feedback on all emails (which would be unrealistic, given that most emails are legitimate and the feedback can be distracting) (Smillie et al., 2013). Kumaraguru et al. (2010) found that embedded

training increased d' and decreased c to reduce overall phishing susceptibility. In a simulation combining SDT and prospect theory, Kaivanto (2014) found that one of the most effective levers for behavioral interventions is increasing the perceived base rate.

In general, warnings tend to reduce c without changing d' (Kumaraguru et al., 2010). Egelman et al. (2008) found that 79% of participants heeded active anti-phishing browser warnings that required acknowledgement before proceeding to the webpage. For passive warnings, users are easily incentivized to ignore security advice to accomplish whatever task is at hand (Christin et al., 2011; Herley, 2009; Herley, 2014). In addition, warnings may be less effective if they do not align with users' mental models (Bravo-Lillo et al., 2011). For example, novice computer users tend to rely on cues related to the "look and feel" of an email or website rather than to more informative technical indicators, such as the URL (Downs et al., 2006). A warning that alters the "look and feel" of a website may be more persuasive (Egelman et al., 2008). Warnings that are distinctive and clearly state that there is a phishing risk, rather than some general problem, are more effective (Carpenter et al., 2013; Egelman et al., 2008).

The vigilance literature suggests that it is not sufficient to simply teach users about phishing risks. Given the low base rate and convoluted payoffs of phishing attacks, users will rationally adopt a bias toward perceiving emails as legitimate. However, institutions can shift the landscape via interventions such as embedded training to help ensure that users appreciate and act on the risk. This thesis bridges

the vigilance and phishing susceptibility literatures to improve the quantification of phishing susceptibility and identify effective behavioral interventions.

3. Quantifying Phishing Susceptibility for Detection and Behavior Decisions

3.1. Introduction

Phishing is among the top cyber attack vectors (Symantec, 2016; Verizon, 2016) threatening individuals, corporations, and critical infrastructure (Wueest, 2014). These attacks seek to trick users into thinking an email or website is legitimate, hoping to convince them to divulge usernames and passwords, or inadvertently install malware by clicking on malicious links or attachments. Depending on the level of deception involved, it can be difficult to screen such messages automatically. As a result, human judgment plays a role in all cybersecurity systems and, by many accounts, is its weakest link (CERT, 2013; Cranor, 2008).

We use signal detection theory (SDT) methods to assess phishing vulnerability by treating phishing detection as a vigilance task (Mackworth, 1948; See et al., 1995; Warm et al., 2008). SDT has been used in a wide variety of contexts, including baggage screening (Wolfe et al., 2013), sexual intent (Farris et al., 2008), medical decision-making (Mohan et al., 2012), environmental risk perception (Dewitt et al., 2015), and phishing detection (Kaivanto, 2014; Kumaraguru et al., 2010; Mayhorn & Nyeste, 2012; Sheng et al., 2010; Welk et al., 2015). By quantifying performance, SDT offers metrics for analyzing system vulnerability, as well as for designing and evaluating interventions to reduce it, such as training, incentives, and task restructuring (Mumpower & McClelland, 2014; Swets et al., 2000). Such

research meets a growing need to integrate human decision-making and perceptual ability into cybersecurity systems (Boyce, 2011; Proctor & Chen, 2015).

The premise of SDT is the need to separate users' *sensitivity* or d' (i.e., their ability to tell whether an email is phishing) from their *response bias* or c (i.e., their tendency to treat an email as phishing) (Macmillan & Creelman, 2004). Accuracy measures such as the number or proportion of successful phishing attacks are incomplete because they ignore other objectives, such as opening legitimate emails promptly. SDT accommodates the inevitable tradeoff between hit rates (*H*, correctly identifying a signal) and false alarm rates (*FA*, incorrectly identifying noise as signals).

The present study demonstrates a procedure for estimating individual users' sensitivity and response bias for phishing, in examining performance on two interrelated tasks: (a) *detection*, deciding whether an email is legitimate and (b) *behavior*, deciding what to do with an email. Unlike many signal detection tasks, where the contingent behavior is straightforward (e.g., rescreening detected bags entails minimal costs for false positives; Wolfe et al., 2007), with phishing, detection and behavior decisions are not uniquely coupled. For example, not falling for a phishing email might reflect discrimination or disinterest. As a result, we study detection and behavior separately in order to assess their respective contributions to vulnerability.

Because behavior has more immediate consequences than detection, we expected greater caution with behavior (Lynn & Barrett, 2014). However, we had no reason to expect differences in sensitivity, unless the more immediate consequences

of the behavior task elicit greater effort, revealing discrimination ability not tapped by detection.

3.1.1. Factors that Influence Signal Detection Estimates

Previous signal detection research has identified a variety of task, individual, and environmental variables that can affect performance (Ballard, 1996). Here, we study behavior as a function of participants' awareness of two such variables: (a) signal base rate (i.e., how frequently the signal appears) and (b) costs for correct and incorrect choices (Coombs, Dawes & Tversky, 1970; MacMillan & Creelman, 2004). These variables have typically had effects consistent with rational decision-making. For example, people are more likely to identify a stimulus as noise for low base-rate events, where it is unlikely to be a signal. Conversely, people are more likely to identify a stimulus as a signal when missing a signal is more costly and a false alarm is less costly (Lynn & Barrett, 2014; Maddox, 2002; Navalpakkam et al., 2009).

The base rate and costs are related to response bias in the following equation, combining Eq 6.4 in Coombs, Dawes & Tversky (1970) and Eq. 2.6 in MacMillan & Creelman (2004):

$$\frac{P(x|s)}{P(x|n)} \geq \frac{1-p}{p}\left[\frac{C_{FA}+C_{TN}}{C_M+C_H}\right] = \beta = e^{cd'}$$

The first term is the likelihood ratio of a stimulus being a signal (s) or noise (n); p is the base rate of the signal; the bracketed term is the *cost ratio*, incorporating the cost of false alarms (FA), true negatives (TN), misses (M), and hits (H); and β is a measure of bias related to c and d' (as seen in the final term). When the likelihood

ratio is greater than β, an observer should treat the stimulus as a signal. Assuming that d' remains constant with changes in task, c should respond to changes in p and the cost ratio (Lynn & Barrett, 2014). We consider both task features in the study design.

3.1.1.1. Signal base rate

Due to the volume of legitimate email traffic and the use of automatic screening programs, phishing emails typically have a low base rate (< 1%) (Symantec, 2016; Verizon, 2016). In the context of baggage screening, Wolfe et al. (2007) describe a prevalence effect, whereby users are biased toward identifying stimuli as noise when there is a low base rate, leading to low hit and false alarm rates. The demands of experimental research typically lead to tasks with artificially high base rates (e.g., Mohan et al., 2012) in order to keep costs down and participants engaged. Participants are, however, typically not told the base rate, leaving it unclear whether they assume a low base rate (as in their lives) or a much higher one due to the experimental context ("they wouldn't ask me to look for phishing emails, if they weren't going to present them fairly often"). They may also infer the base rate based on their intuitions regarding whether experimental stimuli are signals or noise (Wolfe et al., 2007). Here, we examine the effects of explicitly informing participants that the phishing base rate is 50%. If participants who receive no notice infer a 50% base rate, then notification should have little effect. If they infer a lower base rate, then their c should be much higher, indicating less caution regarding attacks.

3.1.1.2. Costs

The consequences of successful phishing can vary widely across domains. The cost of failed detection could be very high, as with critical infrastructure (e.g., an electrical grid blackout), or fairly low, as with a personal laptop (e.g., an annoying virus). Often, users have little direct guidance about those consequences, beyond general cautionary messages (Carpenter, Zhu & Kolimi, 2014). They may also have limited opportunities to learn from experience, as when time separates the attack and its damage or when users provide portals to attack distant targets. Incentives may also be misaligned, as when individuals bear the costs of avoidance actions, while the benefits accrue to the system (e.g., Herley (2009) discusses rational rejection of security advice).

In detection tasks without a clear payoff structure, participants typically try to maximize accuracy (Maddox, 2002), which would produce c = 0 (at a 50% base rate). However, phishing avoidance is an everyday task. In order to capture participants' natural cost expectations, as best we could, we did not impose a cost structure, but compared c for the detection and behavior tasks, expecting less caution for the former, with its reduced costs. Within each task, we expected individual participants' c values to be correlated with their judgments of the consequences of falling for a phishing attack. For the participants notified of the 50% base rate, we assume that β equals the cost ratio. If the base rate notification condition has no effect, we can make the same assumption for the participants without the notice. If the costs of hits (correctly identifying phishing emails) and true negatives (correctly identifying legitimate emails) are minimal, then $\beta > 1$ (and

hence c > 0), implies a cost ratio with lower costs for misses and greater costs for false alarms. Thus, participants who judge the consequences of misses to be worse should have $\beta < 1$ and a negative (or more cautious) c.

3.1.2. Factors that Influence Phishing Susceptibility

Individuals' performance reflects both their ability and how well they apply it. In order to disrupt that application, attackers choose cues designed to evoke heuristic thinking and reduce systematic processing. For recipients who stop to examine messages, and possess requisite knowledge or experience, potentially useful cues include the sender, embedded URLs, grammar, spelling, sense of urgency, and subject line. Studies have, indeed, found less susceptibility among individuals who pay greater attention to message cues, invest more cognitive effort, have more knowledge and experience, and are more suspicious (Luo et al., 2013; Mayhorn & Nyeste, 2012; Pattinson et al., 2012; Sheng et al., 2010; Vishwanath et al., 2011; Wang et al., 2012; Welk et al., 2015; Wright et al., 2009; Wright & Marett, 2010). Rather than manipulate message features in order to determine participants' sensitivity to them, we use naturalistic stimuli, meant to capture everyday co-variation among the cues. We assess participants' overall feeling for their discrimination ability by eliciting their confidence in their judgment, expecting more confident participants to be more knowledgeable, although not perfectly calibrated (Dhamija et al., 2006; Lichtenstein & Fischhoff, 1980; Sheng et al, 2007). We also use a measure of dispositional suspiciousness, expecting those higher on that trait to perceive worse consequences and be more cautious, but not to differ in their discrimination ability.

3.1.3. Aim of Study

We (1) demonstrate an approach applying SDT to phishing detection, (2) with two interrelated tasks, detection and behavior in response to phishing; and (3) manipulating three task variables: (a) which task comes first, detection or behavior (Experiment 1); (b) whether participants perform both tasks (Experiment 1) or just one (Experiment 2) and (c) whether participants are told, or must infer, the base rate of phishing messages. For each stimulus, we measure participants' (a) confidence, (b) judgments of consequences, and (c) response time.

3.2. Method

3.2.1. Sample

We recruited participants from U.S. Amazon Mechanical Turk (mTurk), a crowd-sourced digital marketplace often used for behavioral research (Paolacci, Chandler & Ipeirotis, 2010). Although mTurk samples are not representative of the general U.S. population, they are more varied than convenience samples like university students (Crump, McDonnell & Gureckis, 2013; Mason & Suri, 2012). mTurk studies often recruit some participants who click through tasks without performing them or perform multiple tasks simultaneously, devoting limited attention to each (Downs et al., 2010). As a result, we use attention checks to measure participants' engagement. This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at Carnegie Mellon University. Informed consent was obtained from each participant.

3.2.2. Design

Following the scenario-based design of Kumaraguru et al. (2010) and Pattinson et al. (2012), participants reviewed emails of a fictitious persona. To reduce participant burden and study costs, phishing emails appear at a high base rate (50%), relative to real-world settings (<1%). We randomly assigned participants to conditions created by crossing three task variables: (1) task order (Experiment 1 only), (2) task type (Experiment 2 only), and (3) notification of base rate.

3.2.3. Stimuli

Participants reviewed emails on behalf of Kelly Harmon, an employee at the fictional Soma Corporation, about whom they received a brief description. Phishing emails were adapted from public archives and descriptions in news articles. Each contained one or more of the following features often associated with phishing: (1) impersonal greeting, (2) suspicious URLs with a deceptive name or IP address, (3) unusual content based on the ostensible sender and subject, (4) requests for urgent action, and (5) grammatical errors or misspellings (Downs et al., 2006). The URL was the most valid cue for identifying a phishing email. Legitimate emails were adapted from personal emails and example emails on the Internet, leading to some phishing cues appearing in legitimate emails (e.g., misspelling). Figure 3-1 shows a phishing email. We randomized the use of personal greetings across all emails, but did not systematically vary other cues. All stimuli mimicked the Gmail format.

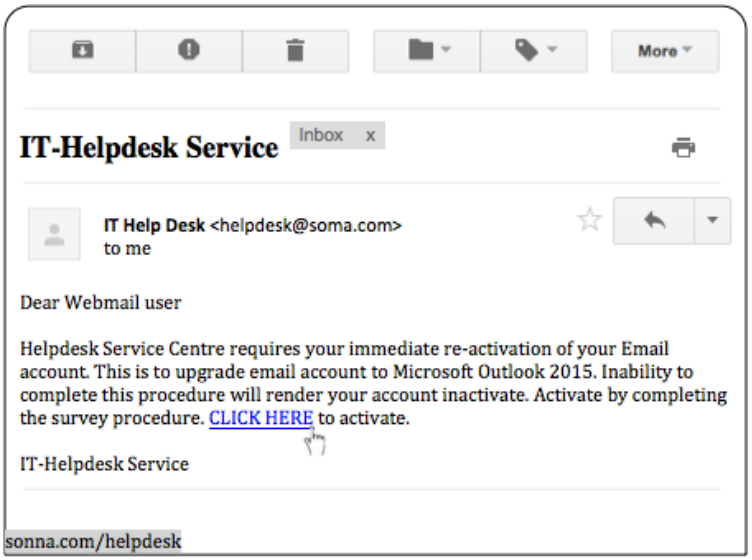

Figure 3-1. A phishing email with all 5 cues.

3.2.4. Measures

Before viewing the stimuli, participants saw one of two messages regarding the base rate: (1) "Approximately half of the emails are phishing emails" or (2) "Phishing emails are included" (*notification of base rate*). In Experiment 1, participants answered the following questions for each email: (1) "Is this a phishing email?" (Yes/No) (*detection*); (2) "What would you do if you received this email?", with multiple-choice options from Sheng et al. (2010) (*behavior*); (3) "How confident are you in your answer?" (50-100%) (*confidence*); and (4) "If this was a phishing email and you fell for it, how bad would the consequences be?" (1= not bad at all; 5=very bad) (*perceived consequences*). Experiment 2 randomly assigned participants to answer either question 1 or 2, rather than both.

To calculate d' and c, the behavior decisions were converted to binary data. Responses of "click link" and "reply," the two actions that could expose users to negative consequences, were interpreted as indicating that participants saw the message as "legitimate"; all other responses were categorized as "phishing."

We included 4 attention checks. At the beginning, two multiple-choice questions asked about the task description: (1) "Where does Kelly Harmon work?" and (2) "What is a phishing email?" Embedded in the task were two email stimuli used as attention checks: (3) "If you are reading this, please answer that this is a phishing email" and (4) "If you are reading this, please answer that this is NOT a phishing email." Many participants saw the "legitimate" stimulus check as suspicious, and identified it as phishing, thereby failing the check (44 for Experiment 1 and 33 for Experiment 2). Therefore, we removed it from the analysis. *Attention* was measured as a binary variable based on the first 3 checks. Rather than removing participants who failed checks, we used attention as a predictor in the regression analyses (below). We found similar results (see Appendix A) when excluding the 10 participants who failed two of three additional attention checks: illogical response (e.g., clicking the link on an email identified as phishing), spending less than 10 seconds on more than one email and $d' < 0$.

We measured the time spent on the phishing information (*phish info time*) and emails (*median time/email*). We used gender, age and education to measure demographic differences. (See Appendix A for details on treatment of these variables.)

3.2.5. Signal Detection Theory Analysis

SDT assumes that both signals (phishing) and noise (legitimate emails) can be represented as distributions of stimuli that vary on the decision variable (here, having properties of phishing emails). The further apart the distributions, the greater the sensitivity or *d'*. The response bias, *c*, reflects how biased users are

toward treating a stimulus as signal or noise. It is measured by how far their decision threshold is from the intersection of the two distributions. A negative response bias ($c < 0$) reflects a tendency to call uncertain stimuli signals. With phishing as the signal, negative values of c reflect a tendency to call uncertain messages phishing, indicating greater aversion to misses (treating phishing messages as legitimate) than to false alarms (treating legitimate messages as phishing).

We estimated the SDT parameters by assuming the signal and noise distributions were Gaussian with equal variance (Lynn & Barrett, 2014). To accommodate cases where participants identified all stimuli correctly or incorrectly, producing hit (H) or false alarm (FA) rates of 0 or 1, a log-linear correction added 0.5 to the number of hits and false alarms and 1 to the number of signals (phishing emails) or noise (legitimate emails) (Hautus, 1995). Thus:

$$H = (\text{hits} + 0.5)/(\text{signals} + 1)$$

$$FA = (\text{false alarms} + 0.5)/(\text{noise} + 1)$$

$$d' = z(H) - z(FA)$$

$$c = -0.5(z(H) + z(FA))$$

3.3. Experiment 1

3.3.1. Procedure

Participants received information about phishing and then evaluated 40 emails. The information was the PhishGuru comic strip from Kumaraguru et al. (2010). It noted that attackers can forge senders and warned, “don’t trust links in an

email." For the email evaluation task, participants examined 19 legitimate emails, 19 phishing emails, and 2 attention check emails. For each email, participants performed the detection and behavior tasks, then assessed their confidence in their judgments and the perceived consequences if the email was phishing. The order of the emails was randomized for each participant. The order of the detection and behavior task was randomized across participants.

3.3.2. Sample

Of the 162 participants who started the experiment, 152 finished. They were paid $5. According to self-reports, 58% were female and 45% had at least a Bachelor's degree. The mean age was 32 years old, with a range from 19 to 59.

Of the 152 participants, 15 failed at least one attention check. For the scenario checks, 3 failed the work question and 9 the phishing question. For the stimuli check, 5 failed the "phishing" version. They spent a minute or two (*Mdn* = 0.95 min, *M* = 3.2 min, *SD* = 11.5 min) on the phishing information and just under a minute per email (*Mdn* = 43 sec, *M* = 52 sec, *SD* = 38 sec), with a median overall time of 40 minutes.

3.3.3. Results & Discussion

3.3.3.1. Phishing detection performance.

We estimated d' and c for the detection and behavior tasks separately, denoted by subscripts D and B, respectively.

Table 3-1 shows aggregate performance. Figure 3-2 shows individual performance. Additional analysis found that d' and c were constant over the course of the experiment (i.e., no learning occurred, see Appendix A for details). We also estimated the area under the curve (AUC) for the individual ROC curves, which is comparable to d', and β, which is a function of d' and c. Closer inspection (detailed in Appendix A) suggests that some participants in both tasks appeared to treat misses and false alarms as equally costly ($\beta = 1$), effectively making accuracy their criterion. In the behavior task, most participants appeared to minimize misses ($\beta_B < 1$). However, their thresholds varied widely for the detection task, with most aiming to minimize false alarms ($\beta_D > 1$). As expected, average perceived consequences was negatively correlated with both β_D, $r(150) = -0.26$, $p = .001$, and β_B, $r(150) = -0.25$, $p = .002$, indicating that participants who perceived worse consequences had lower implicit cost ratios. Participants with a higher β_D also had higher β_B, $r(150) = 0.36$, $p < .001$.

Detection task. Participants' mean sensitivity ($d'_D = 0.96$) indicated modest detection ability. Their mean response bias ($c_D = 0.32$) meant that they had to be somewhat suspicious before treating a message as phishing. These parameters are equivalent to a miss rate of 44% and a false alarm rate of 24% -- both of which would be punishingly high for many computer systems. As seen in Figure 3-2a, both parameters varied considerably across participants. Some had $d'_D < 0$, meaning they consistently misidentified stimuli. Most had positive c_D values. Such variability suggests that a system's vulnerability might be very different depending on whether

it was determined primarily by the average user, the worst user (in terms of d' or c), or the best user (as a sentinel for problems).

Behavior task. When asked how they would respond to each email, participants demonstrated lower sensitivity (d' = 0.39), along with a bias toward not clicking on links (c = -0.54). This combination is equivalent to a miss rate of 28% and a false alarm rate of 61%, also punishingly high for many systems. Figure 3-2b shows the variability in individual performance. Performance on the two tasks was correlated. Participants with a high d' in the detection task tended to also have a higher d' for the behavior task, $r(150) = 0.61$, $p < .001$. The same was true for response bias, $r(150) = 0.66$, $p < .001$.

Figure 3-3a shows responses on the behavior task, based on whether the participant judged a message to be phishing or legitimate in the detection task. Although participants sometimes acted cautiously with messages that they perceived as legitimate (e.g., checking the link or sender), they rarely chose to "click link or open attachment" for emails they perceived as phishing. Figure 3-3b shows these actions as a function of whether the messages were actually legitimate or phishing. Given participants' imperfect detection ability, behaviors consistent with their beliefs sometimes led to inappropriate actions. Thus, despite the bias toward not clicking on links revealed in c_B, participants still succumbed to many phishing attacks. They knew what to do with legitimate and phishing emails, just not which they were facing.

Table 3-1. SDT Performance Parameter Estimates

	Detection Task		Behavior Task		
	Experiment 1 M (SD)	Experiment 2 M (SD)	Experiment 1 M (SD)	Experiment 2 M (SD)	Typical Range
d'	0.96 (0.64)	0.98 (0.80)	0.39 (0.50)	0.41 (0.54)	0 to 4
c	0.32 (0.46)	0.30 (0.44)	-0.54 (0.66)	-0.75 (0.73)	-2 to 2
AUC	0.71 (0.12)	0.70 (0.14)	0.66 (0.12)	0.66 (0.12)	0.5 to 1
β	1.59 (1.13)	1.73 (1.49)	0.88 (0.56)	0.95 (0.43)	0 to 10
H	0.56 (0.19)	0.57 (0.19)	0.72 (0.21)	0.79 (0.16)*	0 to 1
FA	0.24 (0.16)	0.25 (0.18)	0.61 (0.21)	0.65 (0.25)	0 to 1
Accuracy	0.67 (0.11)	0.67 (0.13)	0.56 (0.08)	0.43 (0.09)***	0 to 1

Note: Significant difference between Experiment 1 and 2 based on two-sided t-test where *p <.05, ***p<.001

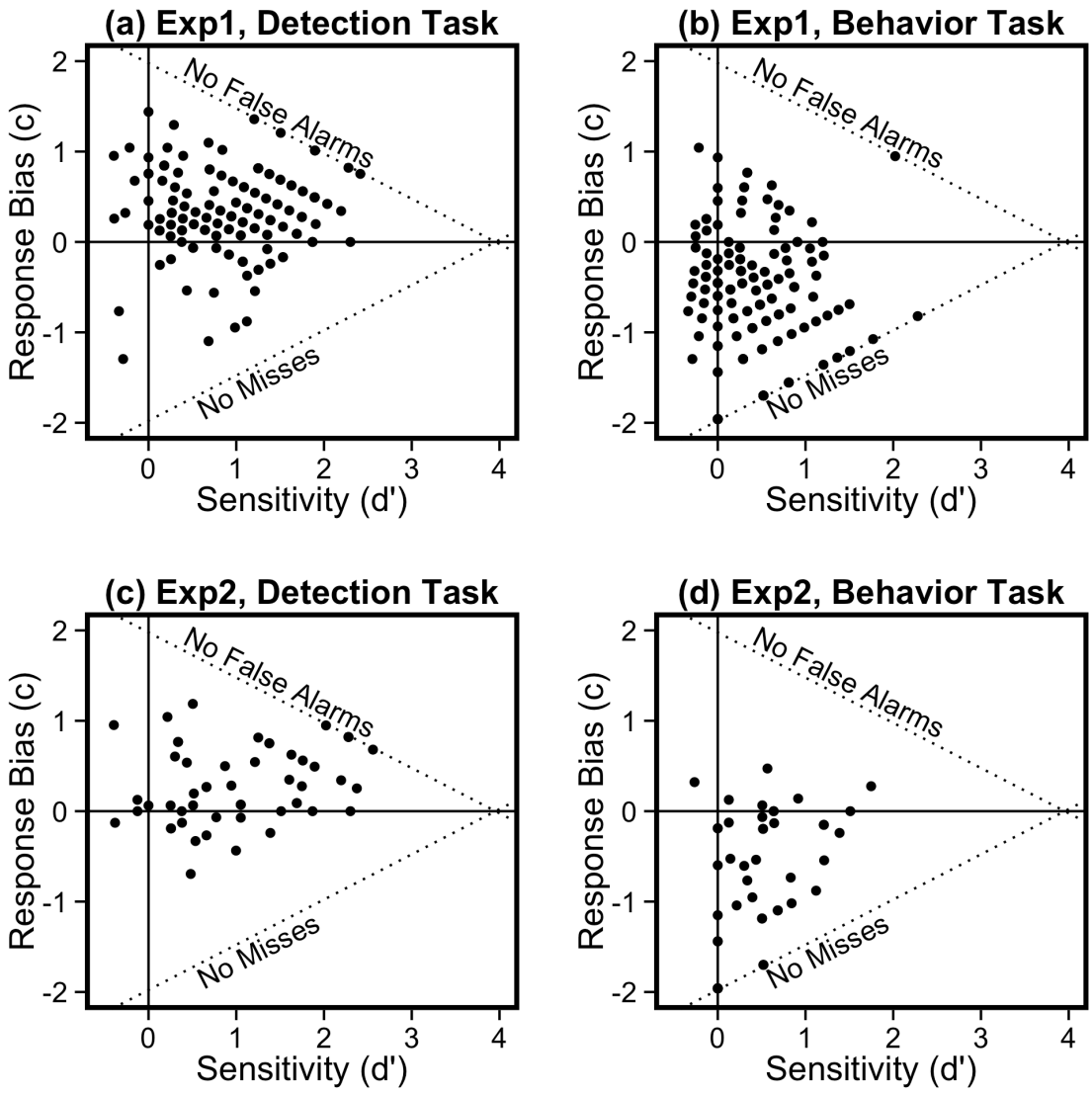

Figure 3-2. Individual variation for detection and behavior tasks in Experiment 1 and 2. The dotted lines denote the mathematical bounds for performance with a false alarm or miss rate of 0%.

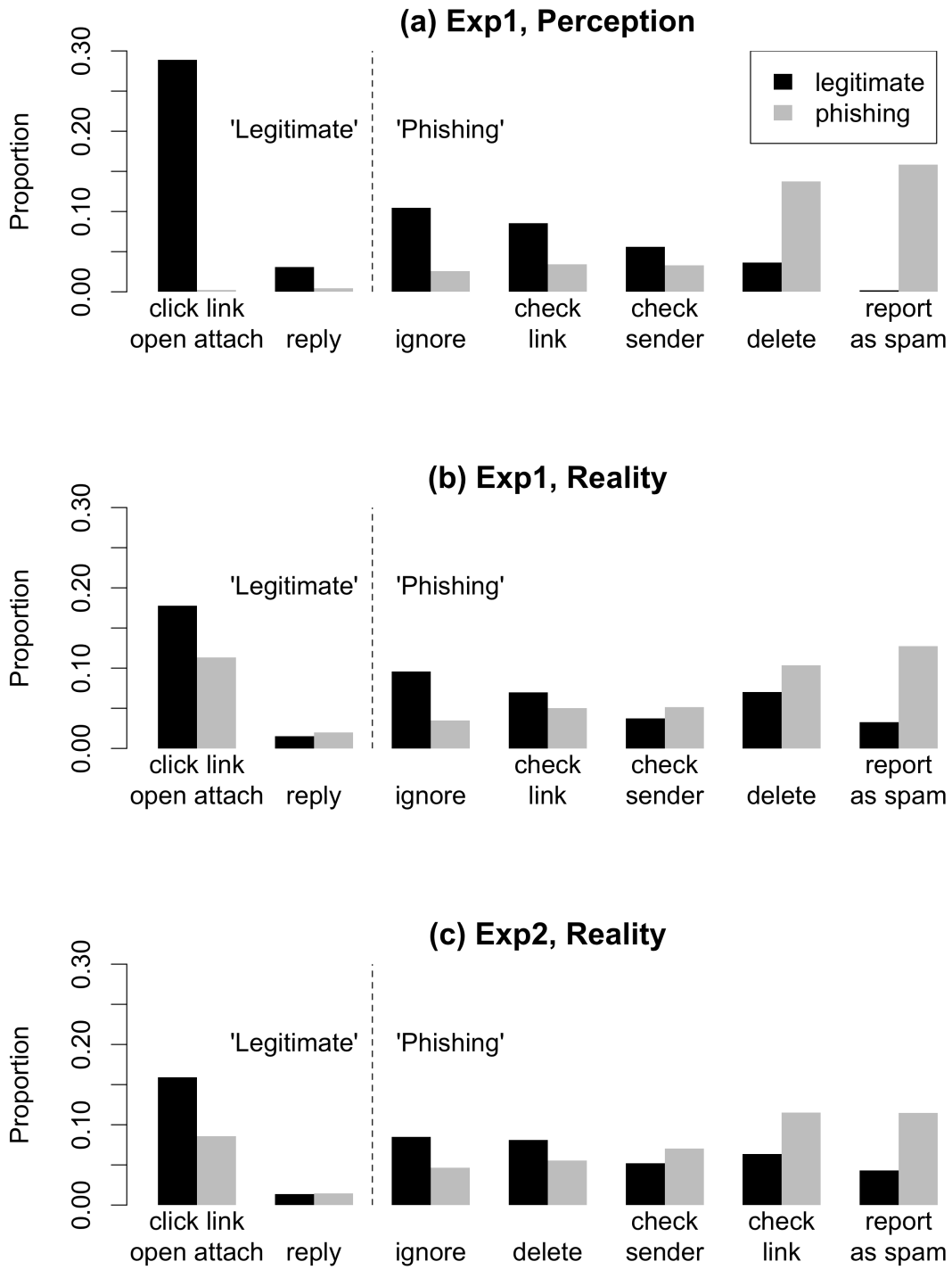

Figure 3-3. Proportion of behavior based on (a) perceived and (b, c) actual type of email.

3.3.3.2. Regression analysis

Table 3-2 and Table 3-3 show multivariate linear regression models predicting individual participants' d' and c between-subjects. Model 1 considers the two between-subjects experimental task variables: (1) notification of base rate and (2) task order. Model 2 adds participants' other responses: attention, phishing information time, median time per email, mean confidence, and mean perceived consequences. Model 3 adds the three demographic measures: age, gender, and college degree. Given the number of statistical tests (11), we use alpha = .01 as the threshold for significance and include tests at the alpha = .05 level, for the reader's convenience.

Model 1: Manipulated between-subject variables. Whether participants performed the detection or the behavior task first did not predict d' or c for either task, nor did whether they received explicit notification of the base rate, *p*'s > .01.

Model 2: Responses to stimuli. Participants who failed the attention checks had lower sensitivity on the detection task, but were no different on the other performance parameters. Thus, users who paid less attention also exhibited lower discrimination ability, but did not differ in how cautiously they acted, given their perceptions. Time spent on the phishing information was not correlated with d' or c, for either task. Participants who spent more time per email were less likely to click on links (i.e., lower c_B), but were no different on the other parameters. The median time spent on each email was uncorrelated to confidence and perceived consequences, *p* > .05.

Table 3-2. Regression models for d' (Experiment 1).

	Detection Task			Behavior Task		
	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables B (SE)	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables B (SE)
Intercept	0.91 (0.09)***	-2.12 (0.67)**	-1.32 (0.98)	0.34 (0.07)***	-1.28 (0.56)*	-0.09 (0.83)
Knowledge of base rate	0.07 (0.10)	0.04 (0.10)	0.02 (0.10)	0.11 (0.08)	0.10 (0.08)	0.10 (0.08)
Task order (detection = 1)	0.02 (0.10)	0.08 (0.10)	0.04 (0.10)	0 (0.08)	-0.01 (0.08)	-0.05 (0.09)
Attention (pass = 1)		0.52 (0.18)**	0.49 (0.18)**		0.15 (0.15)	0.12 (0.15)
log(Phish info time)		0.05 (0.04)	0.05 (0.04)		-0.03 (0.04)	-0.03 (0.03)
Median time/email		0.40 (0.22)	0.48 (0.23)*		0.04 (0.18)	0.17 (0.19)
Average confidence		2.45 (0.65)***	2.23 (0.67)**		1.34 (0.55)*	1.11 (0.57)
Average perceived consequences		0.07 (0.08)	0.08 (0.08)		0.09 (0.06)	0.11 (0.06)
log(Age)			-0.22 (0.21)			-0.33 (0.17)
Gender (male = 1)			0.11 (0.10)			0.06 (0.09)
College (college degree = 1)			0.19 (0.10)			0.10 (0.09)
N	152	142	142	152	142	142
Adjusted R^2	-0.01	0.15	0.16	0	0.03	0.05
F	0.24	4.43***	3.71***	0.84	1.59	1.68

Notes: *p<.05 **p<.01 ***p<.001
Confidence was evaluated from 0.5-1 and perceived consequences were evaluated from 1-5.

Table 3-3. Regression models for c (Experiment 1).

	Detection Task			Behavior Task		
	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables B (SE)	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables B (SE)
Intercept	0.32 (0.07)***	-0.43 (0.47)	0.06 (0.70)	-0.67 (0.10)***	-0.64 (0.59)	0.10 (0.87)
Knowledge of base rate	0.03 (0.07)	0.01 (0.07)	0.01 (0.07)	0.16 (0.11)	0.12 (0.09)	0.13 (0.09)
Task order (detection = 1)	-0.04 (0.07)	-0.01 (0.07)	-0.01 (0.07)	0.12 (0.11)	0.11 (0.09)	0.11 (0.09)
Attention (pass = 1)		0.08 (0.13)	0.08 (0.13)		-0.19 (0.16)	-0.19 (0.16)
log(Phish info time)		0.01 (0.03)	0.01 (0.03)		0 (0.04)	0.01 (0.04)
Median time/email		0.03 (0.15)	0.10 (0.16)		-0.79 (0.19)***	-0.70 (0.20)***
Average confidence		1.68 (0.46)***	1.81 (0.48)***		2.34 (0.58)***	2.38 (0.59)***
Average perceived consequences		-0.24 (0.05)***	-0.24 (0.05)***		-0.43 (0.07)***	-0.42 (0.07)***
log(Age)			-0.17 (0.15)			-0.22 (0.18)
Gender (male = 1)			-0.13 (0.07)			-0.14 (0.09)
College (college degree = 1)			0.02 (0.07)			-0.13 (0.09)
N	152	142	142	152	142	142
Adjusted R^2	0	0.18	0.18	0.01	0.37	0.39
F	0.25	5.34***	4.16***	1.57	13.02***	9.85***

Notes: *p<.05 **p<.01 ***p<.001
Confidence was evaluated from 0.5-1 and perceived consequences were evaluated from 1-5.

For the detection task, participants' sensitivity was positively correlated with their confidence, consistent with having some metacognitive ability (i.e., knowing how much they know). Participants who were more likely to treat emails as legitimate (i.e., higher c_D) also tended to be more confident. Participants who saw more severe consequences were less likely to identify emails as legitimate, but had no different sensitivity. For the behavior task, participants who were more likely to click on links (i.e., higher c_B) tended to be more confident and perceive fewer consequences. We observed no differences in terms of sensitivity.

Model 3: Demographics. No demographic variable was a significant predictor of d' or c, for either task, $p > .01$.

For both tasks, d' and c were unrelated to whether participants were notified of the base rate or which task they completed first. Notification may have had no effect because participants who received no notice assumed a base rate close to 50% (because it was an experiment), or because those who received notice did not (or could not) incorporate the stated base rate in their responses given that there was no feedback (Goodie & Fantino, 1999; Newell & Rakow, 2007). Task order might have had no effect because, once participants performed both tasks on a few stimuli, the two merged in their minds. Experiment 2 examines this possibility, as well as replicating the study as a whole, by having each participant perform just one task.

3.4. Experiment 2

3.4.1. Procedure

Experiment 2 repeats the procedure of Experiment 1, except that participants were randomly assigned to perform either the detection or the behavior task.

3.4.2. Sample

One hundred participants completed the online experiment, with 52 performing the detection task and 48 the behavior task. Participants who had completed Experiment 1 were not eligible for Experiment 2 (and were screened using mTurk qualifications). They were paid $5. The median time spent was 30 minutes. According to self-reports, 48% were female and 40% had at least a Bachelor's degree. The mean age was 33 years old, with a range of 19 to 60.

Of the 100 participants, 9 failed at least one attention check. For the scenario checks, 1 participant failed the work question and 4 the phishing question. Four failed the stimulus check. Presumably because participants only completed one task, the median time per email was lower (*Mdn* = 29 sec, *M* = 43 sec, *SD* = 49 sec), $t(183) = 2.87$, $p = .005$. There was no significant difference in time spent on the phishing information, $p > .05$.

3.4.3. Results & Discussion

In Experiment 2, participants explicitly performed only one of the two tasks. As seen in

Table 3-1 and Figure 3-2, performance was remarkably similar to Experiment 1, where participants performed both. Two-sided t-tests found no significant differences ($p > .05$) between the studies, in sensitivity, response bias, confidence, or perceived consequences. Appendix A provides additional detail.

One possible explanation for the similarity of the results in the two experiments is that people implicitly make a detection decision when making a behavioral choice, and vice versa. As a result, the second task is there implicitly, even when not performed explicitly. If so, then the similarity of the results suggests the robustness of performance on these tasks, which was also unaffected by the order in which they were performed and whether the base-rate was stated. The few differences between the experiments, reported in Appendix A, were in whether coefficients in the regressions were above or below statistical significance (with the signs being consistent).

3.5. General Discussion

SDT disentangles and quantifies sensitivity and response bias. Here, we apply it to distinguishing phishing emails from legitimate ones, looking separately at detection (is this message phishing?) and behavior (how will you respond to it?), building on previous research (Kumaraguru et al, 2010; Pattinson et al., 2012; Sheng et al., 2010; Vishwanath et al., 2011; Wright & Marett, 2010). After reviewing phishing information, participants evaluated 40 email messages on behalf of a fictitious recipient. For each message, they expressed their confidence in their evaluation and rated the severity of the consequences if the email was phishing. Experimental manipulations varied whether the detection and behavior tasks were

performed together or separately, which was done first (when together), and whether the 50% base rate of phishing messages was stated explicitly.

Our results suggest four primary findings. First, participants' behavior almost always reflected appropriate or cautious actions, given their detection beliefs (Figure 3-3). However, their imperfect detection ability meant that such conditionally appropriate behavior still allowed many successful phishing attacks. Thus, it appears that users have learned what to do about phishing, but not when to do it.

Second, the two tasks, deciding whether a message is legitimate and what to do about it, are naturally intertwined. In Experiment 1, performance on the two tasks was correlated, such that participants who had a higher d' for one also had higher d' for the other. Moreover, performance was the same, whichever task was completed first, suggesting that the two could not be separated. Experiment 2 found similar performance with participants who explicitly performed just one of the tasks. Given how intertwined the two tasks seem to be, interventions that address one might naturally address the other. An intervention that succeeded in separating them might improve detection, by focusing users on that task before moving on to behavior, and improve behavior, by allowing time to reflect on the limits to their detection ability. However, as Herley (2009, 2014) observed, slowing the process degrades the user experience, hence might be rejected, even if that is just what users need.

Third, the differences between c_D and c_B suggest that participants used different decision strategies for the two tasks. SDT research has found that

participants' response bias (c) is sensitive to both the base rate and the costs of correct and incorrect choices. The present results suggest that all participants assumed roughly the same (50%) base rate. Stating that rate explicitly made no difference in either experiment, nor was there evidence of learning over the course of the experiment. Therefore, differences in c can be attributed to differences in perceived costs. Although the experiment imposed no actual costs, participants might reasonably have imported cost expectations from their everyday lives.

Responses to the detection task indicated that most participants treated false alarms as more costly than misses ($\beta > 1$), whereas the ratio was reversed for the behavior tasks ($\beta < 1$). Wickelgren (1977) shows how, even when payoffs are clear, people may lack the feedback needed to estimate how well they are achieving their desired tradeoffs. Thus, our estimates of response bias represent the tradeoffs that participants achieved, and not necessarily those that they intended. To the extent that these estimates capture participants' actual preferences, they suggest users engage in relatively lax screening for detection, in contrast to more rigorous evaluation for behavior.

Fourth, individual performance varies widely, for both d' and c. In the regression analyses, the most consistent predictors were participants' confidence in their ability and perception of the consequences. Confidence was strongly related to d' for the detection task, more weakly for the behavior task – consistent with the common result that confidence is positively, but imperfectly, correlated with knowledge (Fischhoff & MacGregor, 1986; Lichtenstein & Fischhoff, 1980; Moore & Healy, 2008; Parker & Stone, 2014). For both tasks, more confident individuals had

higher values of c, hence were more willing to treat messages as legitimate. Participants who saw greater consequences had lower values of c, hence were less willing to treat messages as legitimate, a result found in other studies of phishing detection (Sheng et al., 2010; Welk et al, 2015; Wright & Marett, 2010). In future research, better measurement of perceived consequences might improve these predictions and clarify the causal relationship between caution and confidence.

Future research using SDT also offers the possibility of assessing the effects of interventions that might affect both d' and c, such as brief training exercises at a high base rate with full feedback (Kaivanto, 2014; Wolfe et al., 2007; 2013), phishing detection games (Kumaraguru et al., 2010; Sheng et al., 2010; Welk et al., 2015), and communicating cost information (Davinson & Sillence, 2010; Hardee; Mayhorn & West, 2006). That research could also examine the effects of targeting users who pose the greatest threat to system performance (Egelman & Peer, 2015), such as those identified here with $d' < 0$ – indicating no detection ability or even systematic confusion.

The patterns observed in these two experiments were robust across three manipulations that could, plausibly, have affected them, namely, notifying participants of the base rate, separating the detection and behavior tasks, and varying their order. Although that robustness increases confidence in these patterns, we would hesitate to generalize the performance estimates observed here beyond the present experimental setting. Speculatively, sensitivity might be better or worse with individuals' personal emails, found in a more familiar context, but also amidst the distractions of everyday life, where monitoring phishing is a

secondary task. Indeed, performance here might be a best-case scenario, with phishing the primary task and a high base rate of signals (Wolfe et al., 2007). Nonetheless, performance here was still imperfect, with evidence suggesting that participants were trying: attention checks, orderly regression results, robustness of replication, and differential responses to the detection and behavior tasks that plausibly reflect real-world sensitivity.

Overall, participants exhibited cautious, informed behavior. However, their detection ability was sufficiently poor that their behavior could imperil computer systems dependent on this human element. Based on these results, two promising places for system operators to focus are helping users to understand the consequences of successful phishing attacks and the validity of the signal sent by their own feelings of confidence.

4. Comparing Phishing Vulnerability in the Lab to Real World Outcomes

4.1. Introduction

Translating human behavior from the laboratory to the real world is complex. When in a laboratory environment, participants know that they are being observed and may shift their behavior to better align with (or perhaps frustrate) the perceived research goals (Orne, 1962). In the real world, those potential pressures are lacking, but other varied, possibly unknown and unmeasured, variables may influence behavior. When measuring human behavior in either context, researchers must establish the validity of their measurement procedures. Three forms of validity are commonly considered, *face validity*, in the sense that a measure looks like the phenomenon that it is claimed to measure; *construct validity*, which assesses whether a measure is correlated to other, theoretically related measures; and *predictive validity*, whether a measure predicts the behavior that it is claimed to measure (Cronbach & Meehl, 1955).

Here, we assess the validity of our experimental measures of phishing detection with users from the Security Behavior Observatory (SBO), a field study gathering detailed data on a community sample of computer users' security habits over time (Forget et al., 2014). Specifically, we (1) assess face validity by examining the generality of our experimental results with this community population (vs. Amazon mTurk); (2) evaluate the construct validity of those results, by comparing them with the Security Behavior Intentions Scale (SeBIS) (Egelman & Peer, 2015); and (3) evaluate the predictive validity of those measures by comparing them to

real-world negative outcomes, as reflected in malicious URLs, malware, and malicious files observed on users' home computers. However, as described below, the relationship between susceptibility to phishing attacks and evidence of the security lapses, as measured by the SBO measures, is not straightforward.

4.1.1. Home Computer Security

Maintaining security on a home computer is difficult. Home users often don't know which security practices are most important (Ion et al., 2015) and may have beliefs that conflict with common security advice (Camp, 2009; Wash, 2010; Wash et al., 2015). Users are expected to keep their system (individual software programs as well as operating system) up to date, avoid suspicious links and attachments (e.g. phishing attacks), choose secure passwords and install security programs (e.g. antivirus). Many struggle to follow all these recommendations, despite best intentions.

At the same time, cyber attacks are becoming more varied and pervasive (Symantec, 2016; Verizon, 2016). For example, phishing attacks are no longer limited to email, but may occur over instant messenger, social media, or text messages (Symantec, 2016). There are products to help protect users. For example, email providers use spam filters, browsers employ blacklists to block malicious websites (Sheng et al., 2009b) and anti-virus programs block and delete malicious files and software. In some cases, this requires user engagement, for example, users must update their anti-virus program. In other cases, such as browser blacklists, users have no control.

As with any threat, negative computer security outcomes are related to exposure as well as vulnerability. Thus, unsophisticated or careless users may escape harm if they use their computers little or avoid dangerous situations – perhaps as a result of recognizing their limits. Conversely, knowledgeable users may ward off a high proportion of attacks, yet still succumb if they use their computers heavily or are a valuable target, subject to particularly effective attacks (such as spear phishing). Research on phishing susceptibility suggests that individuals with higher computer literacy are less susceptible to phishing attacks (Sheng et al., 2010; Wright & Marrett, 2010), which may be strongly enough correlated to frequency of computer use (Appel, 2011) to overcome the increased opportunities to succumb. However, research on the SBO data suggests that security engagement, as measured in user interviews, is not a good predictor of security outcomes (Forget et al., 2016).

In the studies that follow, we first ask whether SBO participants perform similarly to mTurk participants on the phishing detection experiment studied in Chapter 3. Next, we compare performance on the experimental tasks with an individual difference measure of security awareness, the Security Behavior Intentions Scale (SeBIS) (Egelman & Peer, 2015). Finally, we assess whether experimental performance correlates with measures of real-world vulnerability available in the SBO, considering as best we can the potential confounds described above.

4.2. Method

4.2.1. Sample

SBO participants were recruited from Pittsburgh-area participant pools and predominantly include retirees and college students. Participants agreed to have the SBO software installed on their personal computers, which collects data on browsing, installed applications, processes, network connections, events, and more. Active SBO participants, who contributed more than a week's worth of data between October 2015 and February 2016, were recruited to participate in the phishing detection experiment. In addition to their monthly compensation for the SBO, each participant received $20 upon completing the phishing detection experiment. For participants who did not start the experiment, we sent 1 reminder. For participants who started, but did not finish the experiment, we sent 2-3 reminders.

Ultimately, we recruited 132 SBO participants to participate in the phishing detection experiment. Of those, 121 started the survey and 98 finished, for a 74% response rate. We excluded 5 participants who had less than 7 days of data in the SBO database. The final sample represents 44% of all the SBO participants (including inactive participants) at that time. As shown in Table 4-1, the SBO sample was older, $t(130) = 4.32$, $p < .001$, and had a higher proportion of college-educated individuals, $t(214) = 3.16$, $p = .002$, than the mTurk sample in Chapter 3. There was no difference in terms of gender, $\alpha = .05$. Older participants tended to be more educated, in part because some of the younger participants were in college (thus had not yet completed their degree), $r(95) = 0.33$, $p = .001$. The SBO sample resembled the SBO population on these variables.

Table 4-1. Comparison of mTurk and SBO demographics.

Variable	mTurk	SBO Sample	All SBO
Female	58%	60%	61%
Bachelors+	45%	63%	58%
Age	32 [19, 59]	41 [19, 81]	46 [19, 87]
N	151	93	213

4.2.2. Phishing Detection Experiment (Laboratory)

We measured phishing detection ability using signal detection theory (SDT) (Chapter 3). Signal detection theory is a method for distinguishing between users' ability to tell the difference between phishing and legitimate emails (sensitivity or d') and bias toward identifying uncertain emails as phishing or legitimate (response bias or c). We followed the scenario-based design of Kumaraguru et al. (2010) and Pattinson et al. (2012), with participants reviewing emails of a fictitious persona. Participants received information about phishing and then evaluated 40 emails. Phishing emails appeared in our lab experiment at a high base rate (50%), relative to real-world settings (<1%), in order to reduce participant burden. The information was the PhishGuru comic strip from Kumaraguru et al. (2010).

We used the same measures described in Chapter 3, repeated here for completeness. Before viewing the emails, participants saw one of two messages regarding the base rate: (1) "Approximately half of the emails are phishing emails" or (2) "Phishing emails are included" (*notification of base rate*). For each email, participants answered the following questions: (1) "Is this a phishing email?" (Yes/No) (*detection*); (2) "What would you do if you received this email?", with multiple-choice options from Sheng et al. (2010) (*behavior*); (3) "How confident are you in your answer?" (50-100%) (*confidence*); and (4) "If this was a phishing email

and you fell for it, how bad would the consequences be?" (1= not bad at all; 5=very bad) (*perceived consequences*). We also measured *attention* (binary measure based on 3 questions: "where does Kelly Harmon work?", "what is a phishing email?", and an email that said "If you are reading this, please answer that this is a phishing email."), the time spent on the phishing information (*phish info time*), and median time spent on each email (*median time/email).* We also collected demographic information on gender, age, and education.

We estimated four phishing detection performance measures: (1) *detection sensitivity (d'_D)*, the ability to tell the difference between phishing and legitimate emails; (2) *detection response bias (c_D)*, bias toward identifying an email as phishing (negative c) or legitimate (positive c); (3) *behavior sensitivity (d'_B)*, the ability to distinguish between when to click on links and when not to; and (4) *behavior response bias (c_B)*, measure of bias toward clicking on links (positive c) or not (negative c).

4.2.3. Security Behavior Intentions Scale

At a separate time in a separate study, 84 participants completed the Security Behavior Intentions Scale (SeBIS) (Egelman & Peer, 2015). The SeBIS has four subscales: device securement, password generation, proactive awareness, and updating. In total, it has 16 statements rated on a Likert scale from 1 (Never) to 5 (Always). The proactive awareness subscale includes 5 statements specifically related to assessing links, such as "when someone sends me a link, I open it without first verifying where it goes" (reverse coded) and "I know what website I'm visiting based on its look and feel, rather than by looking at the URL bar" (reverse coded).

Low scores on the proactive awareness subscale suggest users do not pay close attention to URLs and were related to impulsivity, risk-taking, and dependence (i.e. relying on other people), which is consistent with the phishing detection literature (Egelman & Peer, 2015). Egelman, Harbach & Peer (2016) also found that performance on the proactive awareness scale was correlated with ability to detect a phishing website in a laboratory environment without priming (i.e. being told that they were being tested on their ability to detect phishing websites). The only way to detect that it was a phishing website was to look at the URL. Although only 22 of 718 participants correctly identified the phishing website, their proactive awareness scores were significantly higher than those of the rest of the sample (Egelman, Harbach & Peer, 2016).

4.2.4. Behavioral Outcomes (Real World)

We expect users who are more susceptible to phishing to experience more negative outcomes in real life. Phishing and malware are increasingly intertwined (Sheng et al., 2009a), so we measured 3 negative outcomes related to phishing susceptibility in the SBO dataset: (1) visits to malicious URLs, (2) installed malware, and (3) presence of malicious files. Malicious URLs were identified using the Google Safe Browsing data set for both the browser (Internet Explorer, Chrome, or Firefox) and network packet data. Due to technical limitations for browser extensions, we were unable to collect data from other popular browsers, such as Microsoft Edge. The network packet data include all HTTP traffic for each webpage, while the browser data only records the webpage URL. The average webpage has approximately 100 http requests for the html, CSS, images, ads, multimedia,

JavaScript, Flash and other files that form a single webpage (HTTP Archive, 2016). In addition, http requests can be made from non-browser applications, such as Spotify music streaming. We identified malware via Should I Remove it? (shouldiremoveit.com) and malicious files via VirusTotal (virustotal.com). Malicious files were identified across the entire machine, while malware was limited to installed applications. We assessed each outcome as a binary variable (i.e. 1 = outcome observed at least once and 0 = no outcome observed), rather than a continuous one (i.e., number of negative outcomes) due to the high number of participants who had no negative outcome (i.e. have never visited a malicious website or have no malware) and the unreliability of some of the count data (Long, 1997).

To assess predictive validity, we performed logistic regression, employing a likelihood ratio test to test the degree to which users' signal detection parameter estimates improved model fit for predicting each outcome. We followed the logistic model construction strategy outlined in Hosmer, Lemeshow & Sturdivant (2013) to identify appropriate behavioral predictors to describe the variation in the outcomes. They recommend identifying potential predictors, performing univariate analysis to identify ones related to the outcomes, and eliminating unrelated variables from the regression analysis. The potential predictors are described in Sections 4.2.4.1 and 4.2.4.2.

4.2.4.1. Browsing predictors

We identified potential predictors related to browsing exposure and vulnerability. We expected users with higher browsing exposure to be more likely to

visit malicious URLs in both the browser and network packet data. We identified 3 variables to describe exposure, which were calculated separately for the browser and network packet data to account for the differing scale. These include counts of (1) *total URLs/day*, (2) *unique URLs/day*, and (3) *domains/day*. Each count per day only included active days. Days where the computer was not used or data were not recorded (e.g. on a vacation) were not included in the count of days. We aimed to measure browsing vulnerability in terms of counts of *clicked email links/day*. We expected users who click on more links in emails to be more likely to visit a malicious URL. We assess this in 2 ways, (1) URL tracking, for URLs that include "mail" or "email" after =, &, or ? (excluding email domains), and (2) source data, where the source URL is an email domain and the destination is not – which does not capture links clicked from an email client, such as Outlook). For the network packet data, we were only able to use the tracking method, because source data were unavailable. In addition, for the browser data, we assessed visits to malicious URLs as a function of browser. Most browsers block access to known malicious URLs, but vary in terms of how long that takes, which may or may not be before a user clicks on the link. Internet Explorer is best at blocking malware, but Chrome tends to block malicious URLs faster (Abrams et al., 2014; Drake et al., 2011; Sheng et al., 2009b).

4.2.4.2. Software predictors

We also identified potential predictors related to software exposure and vulnerability. We expected users with higher software exposure to be more likely to have malware and malicious files. We measured software exposure as a count of

total software, excluding updates, installers, and language packages. We also aimed to measure software vulnerability with 3 variables, (1) *delayed software updates*, (2) *days since Windows update*, and (3) *third-party anti-virus (AV)*. Delayed software updates is a count of outdated versions of popular software including Adobe Flash, Adobe Reader, Java, Internet Explorer, Chrome, and Firefox and ranges from 0 to 6. A program was considered outdated if the user was not using the latest version the day after it was released. Days since Windows update is a count of days since a Windows update was installed. This measure does not capture why users waited to install updates (e.g. users who actively delayed updates vs. those had not been prompted because their computer had been off).

For third-party AV, we assigned a binary variable where 1 = running AV and 0 = no AV. An AV program was considered "running" if it was in use for > 7 days, updating without update errors, and scanning. In some cases, it was impossible to know if an AV program met all of these criteria because the data were not logged or the log was not informative. In those cases, we used the available subset of these criteria. Thus, we assumed the AV was running unless there was evidence otherwise. Table 4-2 summarizes the AV status of the users in the sample. We were able to examine the logs for McAffee, Malwarebytes, Webroot, Avast, Norton, Kaspersky, and AVG to assess time. The median was 168 days in use (M = 221, SD = 237). We were unable to assess updating for Avast and unable to assess scanning for McAfee, Avast, and AVG.

Table 4-2. Number of users with each AV and AV status.

Anti-Virus Program	Installed	Log	Updating	Update Errors	Scanning	Detections
Windows Defender	NA	79	79	48	57	20
McAfee	27	48	16	3	Unclear	Unclear
Malwarebytes	24	23	23	9	18	2
Webroot	9	9	9	2	9	6
Avast	7	9	Unclear	Unclear	Unclear	Unclear
Norton	18	2	2	2	2	2
Kaspersky	6	2	2	2	2	1
AVG	2	1	1	1	Unclear	Unclear
Verizon Internet Security Suite	4	NA	NA	NA	NA	NA
LiveUpdate	2	NA	NA	NA	NA	NA
Avira	2	NA	NA	NA	NA	NA
PCKeeper	2	NA	NA	NA	NA	NA
Trend Micro Titanium Internet Security	2	NA	NA	NA	NA	NA
Optimo System Security Suite	1	NA	NA	NA	NA	NA
STOPzilla	1	NA	NA	NA	NA	NA
Zemana AntiMalware	1	NA	NA	NA	NA	NA
No Third-Party AV	19	NA	NA	NA	NA	NA

4.2.5. Study Design

In this study, we (1) replicated experimental results from Chapter 3 in a community population, (2) evaluated construct validity via correlation with SeBIS, and (3) evaluated predictive validity using data from the SBO (as just described). For each outcome, we used a likelihood ratio test to compare the goodness of fit for logistic regression models with and without each of the SDT measures. The likelihood ratio test is the most powerful test of the null hypothesis that the SDT measure does not increase the likelihood of the data given the SDT measure.

To reduce bias and increase transparency, we preregistered the logistic regression models (without SDT measures) at the Open Science Framework (see Appendix B.1) before combining the SBO and experimental data (Miguel et al., 2014;

Nosek & Lakens, 2014). The analysis reported here differs from the proposed analysis reported in the preregistration due to acquiring more SBO data in the interim. Once we began the analysis, and saw the structure of the data, we were able to improve it by (1) eliminating repetitive measures (e.g. counts of unique domains and counts of social media domains), (2) implementing an automated process for identifying malware, rather than relying on manually coded items, and (3) adding malicious files as an outcome variable. We believe that these refinements were implied by the registered analysis plan, although not all were stated explicitly.

4.3. Results

4.3.1. Comparison of Experimental Results

Of the 93 SBO participants, 16 failed at least 1 of the 3 attention checks. Users who failed the attention checks were not excluded from the sample because attention was not a significant predictor of performance in the regression analysis. Proportionally, slightly more SBO participants failed the attention checks than did mTurk participants (17% vs. 10%), $\chi^2(1) = 3.85$, $p = .05$. The median time to complete the experiment was 47 minutes, including breaks (M = 59 min, SD = 2,400 min). SBO participants spent more time per email, SBO = 0.94 minutes (M = 1.13, SD = 0.72) vs. $mTurk$ = 0.48 minutes (M = 0.53, SD = 0.24), Z = 11850, $p < .001$ in a Wilcoxon signed-rank test. Within the SBO sample, older participants spent more time per email, $r(90) = 0.46$, $p < .001$. SBO participants spent less time on the phishing information, SBO = 0.74 minutes (M = 1.16, SD = 1.79) vs. $mTurk$ = 0.95 minutes (M = 3.17, SD = 11.51), Z = 5018, $p = .02$ in a Wilcoxon signed-rank test.

There were no significant differences between the mTurk and SBO samples on any performance parameters, on either the detection or the behavior task, $\alpha = .05$. Table 4-3 shows the mean statistics for the signal detection parameters and for accuracy. Figure 4-1 shows the distribution of d' and c. There was no evidence of learning over the course of the experiment, as d' and c were equal when calculated separately for the first and second half of the emails.

We also replicated the regression analysis to determine which factors predict phishing detection performance. Figure 4-2 plots the regression coefficients for each predictor for both the community (SBO) and mTurk samples (with full statistics in the Appendix B.2). In general, the community sample's coefficients have larger confidence intervals, due to the lower sample size), but overlap with the mTurk coefficients. There are 2 primary differences between the community and mTurk samples. First, confidence was not a significant predictor of c for the SBO sample, even though there was no difference in mean confidence in the two samples, *M* = 0.86 (*SD* = 0.08) for SBO and mTurk, alpha = .05. Second, age and college have a bigger effect in the SBO sample, perhaps due to the higher variance.

Table 4-3. SDT Performance Parameter Estimates.

	Detection Task		Behavior Task		
	mTurk M (SD)	SBO M (SD)	mTurk M (SD)	SBO M (SD)	Typical Range
Sensitivity (d')	0.96 (0.64)	0.96 (0.66)	0.39 (0.50)	0.42 (0.52)	0 to 4
Response bias (c)	0.32 (0.46)	0.20 (0.51)	-0.54 (0.66)	-0.62 (0.57)	-2 to 2
Accuracy	0.67 (0.11)	0.67 (0.11)	0.56 (0.08)	0.57 (0.09)	0 to 1

Note: No significant difference between mTurk and SBO based on t-test, $\alpha = .05$. Reported mTurk results are from Chapter 3.

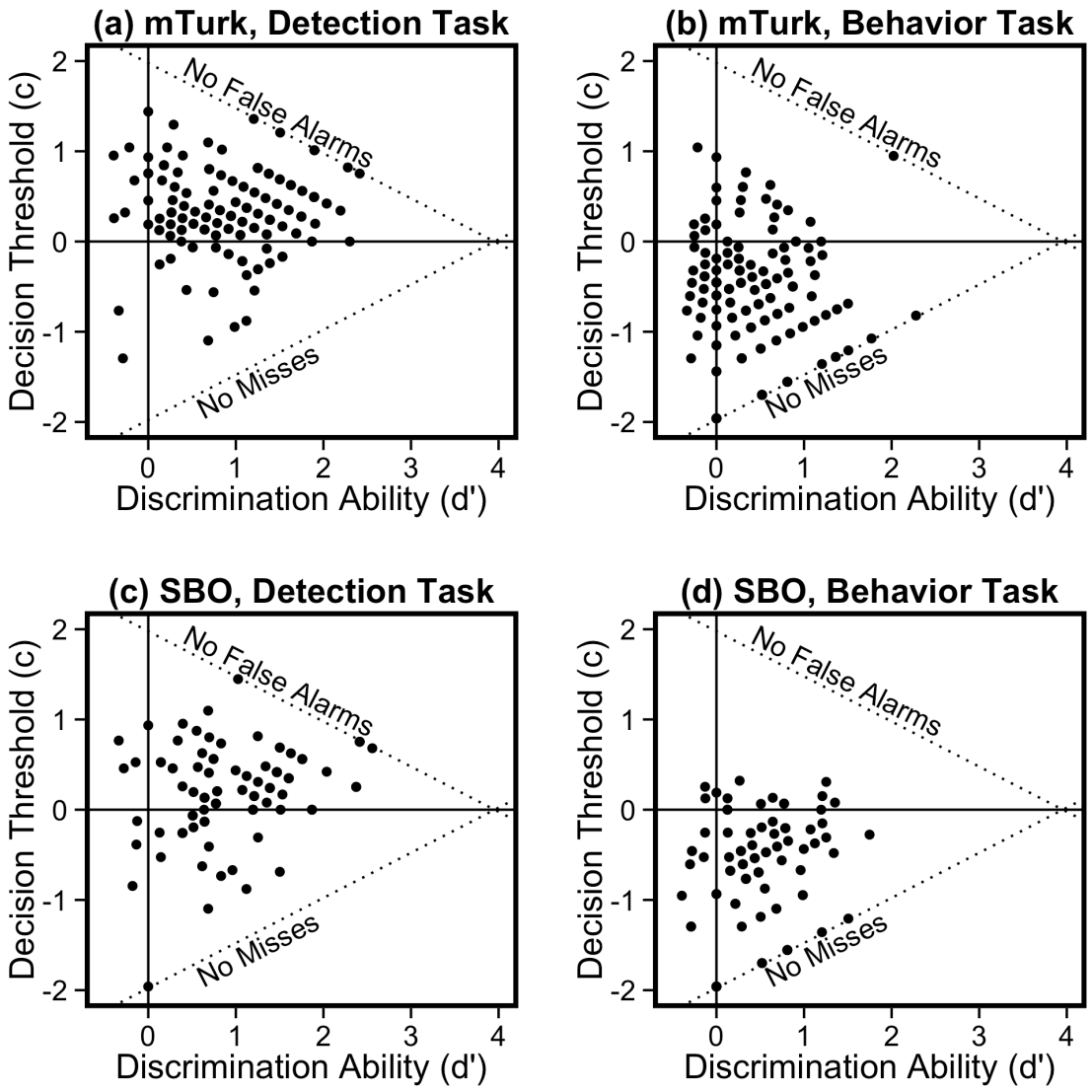

Figure 4-1. Plot of d' vs. c for each task and sample.

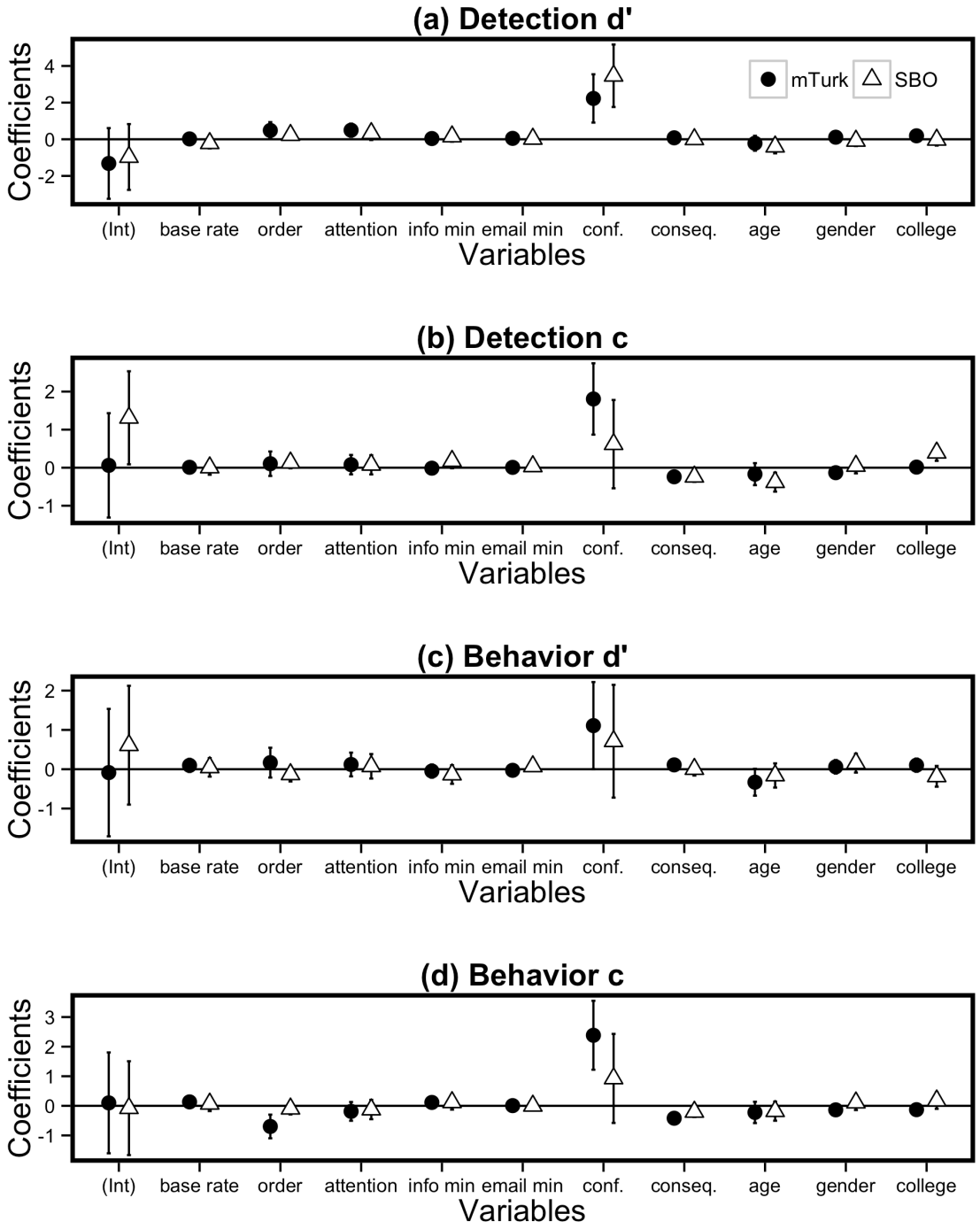

Figure 4-2. Comparison of regression coefficients with 95% confidence intervals (CI) for (a) detection d', (b) detection c, (c) behavior d', and (d) behavior c. Results are reported in a table in Appendix B.2.

4.3.3. Construct Validity

We assessed construct validity via the correlation between the detection and behavior SDT parameters with an existing scale for measuring the same construct: the awareness subscale of SeBIS. Only one of the four SDT parameters, c on the behavior task, was correlated with the proactive awareness subscale, $r(82) = -0.29$, $p = .008$. None of the other SDT parameters had a correlation greater than 0.20 (see Appendix B.2).

4.3.4. Predictive Validity

For ease of explication, we report tests of predictive validity for the behavior task. Results for the detection task are included in Appendix B.2. In general, the findings for the detection task were the same as the behavior task and any differences are noted in the text.

4.3.4.1. Browsing model

Browser sensors. Browser data were available for 86 of the SBO users. Most used Internet Explorer (66/86 = 77%), followed by Chrome (29/86 = 34%) and Firefox (12/86 = 14%). Some used multiple browsers, so the percentages do not sum to 100%. In total, 9 (10%) had visited a malicious URL. Proportionally, more Firefox users had visited malicious URLs (3/12 = 25%) than Chrome (4/29 = 14%) or Internet Explorer users (2/66 = 3%).

Table 4-4 shows descriptive statistics for variables measured in the browser data. Following Homser et al. (2013), we performed univariate analysis between each of these variables and whether users had visited a malicious URL, as indicated

by the browser data. Among these potential covariates, only domains/day was related to the outcome variable, visits to malicious URLs. It was, therefore, included in the regression model using a log transformation to normalize the observations (see Appendix B.2 for detail). Table 4-5 shows the regression analysis for the browser data, predicting visits to malicious URLs. Log(domains/day) was the only significant predictor. Thus, users who visit more domains are more likely to have visited a malicious URL. As seen in the likelihood ratio tests for models 2-4 in Table 4-4, users' signal detection parameter estimates do not improve the model fit, singly or together.

Network packet sensor. We also assessed visits to malicious URLs in the network packet data. As seen in Table 4-4, there is much more network packet data than browser data, which were available for 92 of the 93 participants. This is because the network packet data include all http traffic, which encompasses all of the files required to load each webpage. There are more active days observed for the network packet data, and http requests can be made from other programs besides a browser (e.g. Spotify music streaming).

For 31 of these 93 users (33%), the network packet data indicates they visited a malicious URL. Table 4-4 summarizes the potential covariates. Univariate analysis suggested that total URLs/day, unique URLs/day, and domains/day were related to having visited a malicious URL at least once. We then computed a factor analysis, which revealed that these covariates loaded on one factor, alpha = 0.79. We called this factor browsing intensity and used a log transformation to normalize it (see Appendix B.2 for details). We then used that factor score in the regression

model and likelihood ratio tests reported in Table 4-6. Model 1 shows that users with higher browsing intensity were more likely to have visited a malicious URL in the network packet data. In addition, there was an effect for gender, whereby men were more likely to visit malicious URLs. In the likelihood ratio test for models 2-4, users' signal detection parameter estimates did not further improve the model fit.

Table 4-4. Descriptive statistics and factor analysis for the browser and network packet sensor covariates.

	Browser Sensor		Network Packet Sensor		
	Median	Mean (SD)	Median	Mean (SD)	Loading
Active Days	40	67 (76)	70	85 (63)	NA
Total URLs/ Active Day	22	56 (90)	1,500	2,600 (3,600)	0.73
Unique URLs/ Active Day	9	23 (32)	670	990 (1,000)	1
Domains/ Active Day	5	5.7 (4.4)	42	52 (37)	0.54
Clicked Email Links/ Active Day (tracking)	0	0.5 (1.0)	0	0.4 (2.0)	-
Clicked Email Links/ Active Day (source)	0	0.8 (2.1)	NA	NA	NA
% of Total Variance					61%
Cronbach's Alpha					0.79

Table 4-5. Logistic regression models and likelihood ratio test (LRT) for browser data and behavior task SDT parameters.

	Model 1	**Model 2**	**Model 3**	**Model 4**
(Int)	-5.91**	-5.82**	-6.47**	-6.43**
	(1.90)	(1.92)	(2.06)	(2.14)
Behavior d'		-0.24		-0.06
		(0.86)		(0.89)
Behavior c			-0.81	-0.80
			(0.72)	(0.74)
log(Domains/day)	1.94*	1.97*	1.92*	1.93*
	(0.76)	(0.77)	(0.76)	(0.77)
Age	0.01	0.01	0.01	0.01
	(0.02)	(0.03)	(0.03)	(0.03)
Male	0.75	0.77	0.90	0.90
	(0.78)	(0.78)	(0.80)	(0.81)
College	-1.11	-1.15	-0.88	-0.89
	(0.91)	(0.92)	(0.93)	(0.95)
X^2 (LRT)	11.76*	0.08	1.28	1.29

Table 4-6. Logistic regression models and likelihood ratio test (LRT) for network packet data and behavior task SDT parameters.

	Model 1	**Model 2**	**Model 3**	**Model 4**
(Int)	-10.57***	-10.53***	-10.57***	-10.53***
	(2.84)	(2.83)	(2.83)	(2.83)
Behavior d'		-0.33		-0.33
		(0.55)		(0.55)
Behavior c			0.09	0.11
			(0.50)	(0.50)
log(Browsing Intensity)	1.36***	1.39***	1.37***	1.39***
	(0.38)	(0.38)	(0.38)	(0.38)
Age	-0.03	-0.03	-0.03	-0.03
	(0.02)	(0.02)	(0.02)	(0.02)
Male	1.48**	1.49**	1.46**	1.47**
	(0.54)	(0.55)	(0.55)	(0.55)
College	0.20	0.17	0.20	0.16
	(0.60)	(0.61)	(0.61)	(0.61)
X^2 (LRT)	28.71***	0.36	0.03	0.41

4.3.4.2. Software model

Malware. Most users had the Windows 10 operating system (53/92 = 58%), followed by Windows 8 (22/92 = 24%), Windows 7 (14/92 = 15%), and Windows Vista (3/92 = 3%). 45 of the 92 (47%) with installed software data had malware. Malware (as defined by the ShouldIRemoveIt.com dataset) was observed on machines across all operating systems, with no obvious pattern. For each operating system, approximately half of the users had malware.

Table 4-7 shows descriptive statistics for all potential software covariates. Univariate analysis revealed that total software and delayed software updates were related to malware. However, the factor analysis suggested that these variables were weakly related (alpha = 0.40). When included in the regression model separately, delayed software updates were insignificant, so it was removed from the model. Total software was normalized using a log transformation. Users who installed more software were more likely to have malware on their machine. As seen in Table 4-8 (model 1), this variable predicted malware. The signal detection parameter estimates (models 2-4) did not improve the model fit.

Malicious files. Most users (84/93 = 90%) have malicious files on their machine (as defined by the VirusTotal dataset). In the regression model, we used the same predictors as the malware model, reported in Table 4-7. The regression model and likelihood ratio tests are reported in Table 4-9. Users who have installed more software, as measured by software load, were significantly more likely to have malicious files on their machine. The signal detection parameter estimates (models 2-4) did not improve the model fit.

Table 4-7. Descriptive statistics and factor analysis for software covariates.

	Median	Mean (SD)	Loading
Total Software	244	342 (316)	0.5
Delayed Software Updates	2	2 (1)	0.5
Anti-Virus (binary)	0	0.34 (0.48)	-
Days Since OS Update	71	59 (34)	-
% of Total Variance			25%
Cronbach's Alpha			0.40

Table 4-8. Logistic regression models and likelihood ratio test (LRT) for malware outcome and behavior task SDT parameters.

	Model 1	Model 2	Model 3	Model 4
(Int)	-5.98***	-5.91***	-5.99***	-5.93***
	(1.68)	(1.71)	(1.68)	(1.71)
Behavior d'		-0.10		-0.09
		(0.46)		(0.46)
Behavior c			-0.07	-0.06
			(0.43)	(0.44)
log(Total Software)	1.00**	1.00**	0.99**	0.99**
	(0.31)	(0.31)	(0.31)	(0.31)
Age	0	0	0	0
	(0.01)	(0.01)	(0.01)	(0.01)
Male	0.05	0.07	0.06	0.07
	(0.48)	(0.48)	(0.48)	(0.48)
College	0.57	0.55	0.57	0.56
	(0.52)	(0.53)	(0.53)	(0.53)
X^2 (LRT)	15.74**	0.05	0.03	0.07

Table 4-9. Logistic regression models and likelihood ratio tests (LRT) for malicious files outcome and behavior task SDT parameters.

	Model 1	**Model 2**	**Model 3**	**Model 4**
(Int)	-6.50	-6.70	-6.36	-6.65
	(3.58)	(3.79)	(3.48)	(3.71)
Behavior d'		-1.78		-1.59
		(1.02)		(1.04)
Behavior c			-1.28	-0.90
			(1.16)	(1.22)
log(Total Software)	2.31**	2.73**	2.17**	2.58**
	(0.79)	(0.89)	(0.76)	(0.87)
Age	-0.04	-0.05	-0.04	-0.05
	(0.03)	(0.03)	(0.03)	(0.03)
Male	-0.72	-0.73	-0.75	-0.64
	(0.89)	(0.93)	(0.91)	(0.94)
College	-0.93	-1.41	-0.89	-1.29
	(1.19)	(1.28)	(1.20)	(1.29)
X^2 (LRT)	22.79***	3.53	1.42	4.12

4.4. Discussion

In this study, we (1) replicated the experimental tasks from Chapter 3 in a community population (participants in the SBO), as a reflection of face validity (having no reason to expect differences); (2) evaluated construct validity using SeBIS; and (3) evaluated predictive validity using real world outcomes observed in the SBO.

Regarding face validity, we found very similar performance in the community sample as in the convenience (mTurk) samples. Community participants tended to take longer than mTurk users to complete the laboratory tasks, but still performed similarly. They spent less time on the informational material about phishing. The community sample was older, on average, and within it, older participants tended to take longer per email and have a lower d' and c for the detection task. This was somewhat unexpected, since mTurk users are generally perceived as more tech-

savvy than a community population. However, these results are consistent with the general finding that mTurk samples perform similarly to community samples on psychological tests (Paolacci, Chandler & Ipeirotis, 2010).

Regarding construct validity, we found that the behavior response bias (c) was related to the construct of proactive security awareness, as represented in the SeBIS subscale. Users with a lower c, hence less inclined to click on links in emails, tended to score higher on that subscale, which measures attention to URLs based on self-reports, rather than observed behavior in an experimental setting. If those self-reports are valid, then users who report paying more attention to the URL should perform better on both the detection and behavior experimental tasks. However, Chapter 3 found that people use somewhat different decision-making strategies on the two tasks. For detection, participants tended to have a positive c in order to reduce the frequency of falsely identifying legitimate emails as phishing. For the behavior task, it was reversed, such that participants tended to have a negative c in order to reduce the frequency of falsely identifying phishing emails as legitimate. The regression analysis reported in Chapter 3 suggest that perceived consequences of phishing attacks had a bigger effect for behavior decisions.

Regarding predictive validity, we found no evidence. The SDT parameters on the experimental task were not significant predictors for any of the 4 real-world outcomes captured in the SBO data: visits to malicious URLS in browser and network packet data, malware, and malicious files. However, the 4 measures of negative experience were sufficiently robust that they could be predicted. All were positively related to active computer use, whether measured by their amount of

web browsing or installed software. Thus, it is unclear why the ability to identify suspicious messages in our experimental task did not translate to an ability to identify similar suspicious messages in real life and thereby avoid negative outcomes. Broadly, there are four potential interpretations of our results: (1) the experimental task does not evoke actual behavior with respect to phishing; (2) the experimental task evokes actual behavior but in an environment that lacks ecological validity, in the sense of differing fundamentally from that experienced by SBO users; (3) the SBO measures are confounded by other aspects of users' complex real-world experience, or (4) the measures are noisy enough not to reveal the underlying correlations.

The possibility that the experiment does not measure phishing susceptibility (1) seems unlikely given that the results of the experiment are in line with other research measuring phishing susceptibility. Moreover, performance on the task shows expected correlations with other variables - better performance is associated with greater knowledge, confidence, and intentions.

The lack of ecological validity seems more plausible. One potentially unrepresentative feature of the experimental task is that it has a 50% base rate of phishing emails, much higher than in everyday life. Wolfe et al. (2007) found that artificially high base rates decrease c, but have no effect on d'. A second feature of the experimental task is explicitly asking participants to evaluate each email as phishing (and explaining what phishing is), thereby priming users to detect them. Research by Parsons et al. (2015) suggests that explicitly mentioning phishing artificially increases d', but has no effect on c. Together, these studies suggest that

our estimates of performance are overall better than what would be expected in real life. However, there is no evidence to suggest that this would influence the relative performance of users. If these aspects of the experimental design influence all users similarly, then the correlations across measures should be preserved. That is, we would not expect users who are bad at detecting phishing in real life to be better at it, in this artificial environmental, than users who are good at detecting phishing in real life.

The complexity of real-world environments (for SBO users, among others) complicates the relationship between individuals' general propensities (which d' and c attempt to measure) and their actual experiences. As seen here, bad experiences (in the sense of visiting suspicious URLs and having malicious files) are strongly related to the amount of exposure (in the sense of browsing intensity and software load). Perhaps individuals' opportunity for trouble swamps their ability (d') or propensity (c) to avoid it. Participants' rate of bad experiences may also be related to the protection afforded by their system and their attractiveness as targets for attackers. That vulnerability is partially determined by users (hence related to their abilities) and partly by others (e.g. browser blacklists). Unfortunately, even with the rich SBO data set, we lacked the complete picture needed to sort out these relationships. The SBO includes some data on browser warnings, but there are very few observations. As described in the methods section, we were unable to measure anti-virus events for all anti-virus software. Some of those programs, particularly free versions, do not record logs. Others have poor documentation. Of those that do provide logs, we observed few detections (see Table 4-2). Given that AV use did not

predict the presence of malicious files and the low number of observed detections compared to observed malicious files, it is possible that few users in the SBO sample are able to effectively operate their AV program.

Finally, the SBO measures are noisy, as would be expected from real-world observation. There are cases where data are missing (e.g. a sensor breaks or is turned off) or ambiguous (multiple people using the same computer). As a partial check on one potential source of noise, we repeated the analysis excluding computers with multiple users, but the results did not change.

Overall, there was weak evidence to support the validity of the behavior task, but the results were largely inconclusive. Validating scales using this type of real world data is complex and it is difficult to discern measurable relationships. In the next section, we provide recommendations for validating scales using this type of novel real-world data.

4.4.1. Recommendations

Based on this work, we have 4 primary recommendations for validating performance tests, like our experimental tasks, using real-world data, like that provided by SBO:

1. To the extent possible, measure performance on tasks that are (a) as directly related to the focal outcome as possible and (b) that rely on human ability without intervening technology.
2. Triangulate using multiple data sources (e.g. assessing both browser and network packet data), with an understanding of their respective strengths

and weaknesses. For example, there is more network packet data, but browser data better reflects URLs that users choose to click.

3. Consider the temporal sequence of events, such as how periods without AV protection affect the risk of acquiring malicious files. The present analysis largely ignored this aspect, which may have contributed to the inconclusive nature of the evidence.
4. Create and register an analysis plan in advance. Document deviations in terms that clarify, as best possible, whether they are in the spirit of the plan, responding to unanticipated data structures. Doing so reduces the risk of capitalizing on chances and missing design features that might have improved data quality and relevance.

5. Benefit-Cost of Improving Human Detection of Phishing Attacks: Fixing the Weakest Links

5.1. Introduction

Most cyber attacks begin with a successful phishing attack via email, or increasingly, social media websites (Symantec Corporation, 2016; Verizon, 2016). Phishing (or social engineering) attacks aim to gather information or trick users into inadvertently installing malware, which allows hackers to access networks. Often, attackers mass-email employees, gathering information from out-of-office replies and bounce notices, as well as whatever information users are tricked into providing. This information can then be used to design attacks, called spear phishing, that use personal information (e.g. known contacts, industry language, victims' names) to design more realistic and persuasive messages. When successful, phishing attacks can provide hackers with wide access to an organization's network. At present, many firms are trying to reduce phishing vulnerability, as evidenced by the market for anti-phishing training and analytics (e.g. PhishMe, ThreatSim, Wombat Security).

When employing such behavioral interventions, organizations want to ensure that they are allocating resources cost-effectively. Previous research has found that users' phishing susceptibility varies widely (Pattinson et al., 2012). Here, we create a risk model to evaluate the relative benefit-cost of interventions for subgroups with varying phishing vulnerability. The model considers (1) the identification of poor detectors, (2) the contribution of poor detectors to overall

system vulnerability, and (3) the benefit-cost of interventions targeting poor detectors.

5.1.1. Modeling Phishing Risk

Cybersecurity risk (R) is conventionally defined as a function of threat (T), vulnerability (V) , and impact (I) (NIST, 2012). In this formulation, impact is the cost of a successful attack in terms of money, reputation, productivity, or safety. The probability (P) of a successful attack is a function of threats and vulnerabilities. Threats include malicious attacks by internal and external actors (e.g. phishing) as well as errors (e.g. accidentially publishing private information). Vulnerabilities are human, organizational, or technical weaknesses that can be exploited by an adversary (e.g. zero-day vulnerability) (Sun et al., 2006; Werlinger et al., 2009). These elements are related symbolically by the following equations:

$$R = I*P$$

$$P = F(T, V)$$

It is typically impossible to estimate the absolute value of R. Most notably, the threat is unknown and perhaps varying – in part, as a function of adversaries' perceptions of the vulnerabilities and targets' responses to them. Generally, an organization cannot control threats, but must rely on legal and political authorities for protection. It can, however, try to reduce the impact of attacks (e.g. through network segmentation or limiting permissions across the network) or its vulnerability (e.g. through behavioral interventions). As a result, formal analyses are most useful for comparing the relative risk of alternative system designs, attempting to reduce vulnerability, while assuming that the threat is constant. The class of

model developed here provides a way to evaluate the relative vulnerability of alternative behavioral interventions. It considers both the variability in user performance, offering the possibility of focusing resources on the poorest detectors, and the variability in the effectiveness of behavioral interventions, for phishing and related tasks. Section 5.1.2 reviews the evidence on both forms of variability, translating it into analytic terms.

5.1.2. Accounting for Human Variability

Managing phishing risks is an example of what human factors (or ergonomics) researchers call vigilance tasks, ones in which individuals must monitor their environment for a specific signal. Mackworth (1948) first studied vigilance in order to determine the optimal watch length for airborne radar operators to maximize accuracy in submarine detection.

Since then, vigilance research has identified task, individual, and environmental variables that can affect performance (Ballard, 1996). Task factors include base rate, payoffs, and similarity of stimuli (Lynn & Barrett, 2014). People are less likely to identify a signal when there is a low base rate, the cost of missing a signal is low, the cost of a mistaking noise for a signal is high, or there is very little difference between the signal and noise (e.g. navigating a dimly lit room). Individual factors include experience, personality, and demographics. People are less likely to identify a signal correctly when they are less experienced, more impulsive, older, or less intelligent (Ballard, 1996). Environmental factors reduce performance by increasing stress, such as uncomfortable ambient conditions (e.g. noise or temperature,) and workload (Ballard, 1996). The wide range of shaping factors

suggests that we should expect variation in performance both within and between users (e.g. even highly trained users might occasionally be distracted and fall for phishing attacks). In the following sections, we discuss human variability in terms of susceptibility to attacks and ability to change via behavioral interventions.

5.1.2.1. Variability in Phishing Susceptibility

Following vigilance research, we conceptualize human phishing vulnerability in signal detection theory (SDT) terms, using sensitivity (d') and response bias (c) (Macmillan & Creelman, 2004). Sensitivity (d') refers to users' ability to distinguish between signal and noise, here, phishing and legitimate emails. Larger values of d' indicate greater discrimination ability. Response bias (c) refers to users' tendency to treat an email as phishing or legitimate, when translating their uncertain beliefs into actions. When c is 0, users show no bias. When c is negative, users are biased toward treating emails as phishing and, when c is positive, users are biased toward treating emails as legitimate.

As with vigilance research, phishing detection research has found that vulnerability (i.e., d' and c) can be influenced by task, individual, and environmental factors (Ballard, 1996; Vishwanath et al., 2011; Wright & Marett, 2010). For task factors, Chapter 3 found that users were less discriminating and more cautious (i.e., lower d' and c) when asked to choose an action (e.g. click the link) rather than simply characterize an email as phishing or not, likely because the perceived consequences (or payoffs) were higher for actions. In addition, users who perceived worse consequences of phishing were more cautious (lower c). Providing information about the base rate of phishing emails in the test set had no effect on

performance. Wolfe et al. (2007) did find sensitivity to base rates, in the context of baggage screening, in studies that manipulated the base rate (rather than just told people about it). Spear phishing could be construed as a form of similarity, which reduces d'.

For individual factors, Wright & Marett (2010) distinguish between experiential and dispositional variables. In terms of experience, users with more computer knowledge tend to be less vulnerable (Pattinson et al, 2012; Sheng et al., 2010; Vishwanath et al., 2011; Wang et al., 2012). In terms of disposition, users who are more impulsive (Pattinson et al., 2012) and trusting (Welk et al., 2015) tend to be more susceptible, whereas those who are risk-averse (Sheng et al., 2010) tend to be less susceptible. These individual factors can interact with demographic factors; for example, women tend to have less computer knowledge and younger people tend to be less risk-averse, both of which may make them more vulnerable (Sheng et al., 2010).

Environmental factors, such as workload and time pressure, increase stress, which may increase vulnerability. For example, users who receive many emails and check their email as a habit, without much conscious effort, are more vulnerable (Vishwanath et al., 2011; Vishwanath, 2015). Similarly, users who multi-task while checking their emails or work under tight time deadlines, encouraging cursory review of emails, might be more vulnerable.

At present, a common way for organizations to evaluate phishing susceptibility is via "embedded training" model: sending fake phishing emails to employees, observing who clicks on the links, and (potentially) providing remedial

treatment (Kumaraguru et al., 2010). An alternative strategy is to use an independent measure of phishing susceptibility to identify users needing extra training or protection. Tests of computer security knowledge or attitudes (e.g. Egelman & Peer, 2015) might guide such targeting. We have also developed a test of phishing susceptibility that system operators might employ (Chapter 3). It characterizes vulnerability in terms of d' and c, thereby providing parameter estimates for risk analyses. The next section summarizes evidence regarding the effectiveness of interventions that might be administered to some (or all) of a system's users based on their performance.

5.1.2.2. Effectiveness of Anti-Phishing Interventions

Vigilance researchers have long been interested in improving the detection of low base-rate phenomena. Typically, these are high consequence events (e.g. diagnosing cancer, detecting an enemy submarine, avoiding phishing links) where the cost of missing an event is high, but it is also impossible to treat every case as an impending disaster (because signals are so infrequent). For example, one cannot tell people that they have cancer based on weak signals, just to ensure that all cases are caught (Welch, Schwartz & Woloshin, 2011). Similarly, it is not realistic to treat a large portion of emails as phishing, as that would interfere with users primary work duties. Given how few emails are phishing, the warnings might, at some point be ignored (Wickens et al., 2009).

In vigilance research, most interventions focus on task or individual factors. For example, in the context of baggage screening for airport security, Wolfe et al. (2007) found that exposing operators to brief bursts of training at a high base rate

with full feedback reduced response bias (c), even after returning to a real world with a low base rate without feedback. This suggests that regularly performing such training might encourage observers to maintain a low c despite the low base rate (Wolfe et al., 2013). With air-traffic control, Bisseret (1981) observed that more experienced controllers had a lower c than new recruits, but there was little difference in terms of d'. Such results suggest that experience can reduce the perceived costs of false alarms and encourage reporting.

For phishing detection, common behavioral interventions include embedded training (feedback on misses), warnings (about known risks), and education (ranging from information to games). Most studies have measured performance in terms of accuracy (i.e., the number of successful attacks in some period of observation). However, accuracy conflates d' and c. Accuracy could be increased through better discrimination or more cautious decision rules. In one of the few studies measuring phishing detection performance in signal detection theory terms, Kumaraguru et al. (2010) found that embedded training increased d' and decreased c. Embedded training is similar to the intervention tested by Wolfe et al. (2007), but includes feedback only on false negatives, cases where phishing attacks are missed.

Interventions that increase attention or effort have sometimes been found to increase d'. For example, Parsons et al. (2015) found that telling users that they were being evaluated for their phishing detection ability increased their d' without changing their c. Wolfe et al. (2013) observed an increase in d' during the high base rate training trials. However, unlike the sustained decrease observed with c, d' returned to the previous value immediately after the training. One possible

explanation is that screeners could not sustain the heightened level of attention that they mustered during the training trials.

Table 5-1 and Figure 5-1 summarize studies of behavioral interventions that reported results in SDT terms. They were identified by using the joint search terms of "signal detection theory" and “behavioral intervention” in Google Scholar, which produced 76 papers. We identified an additional 65 papers using the joint search terms “signal detection theory,” “phishing,” and “experiment.” We then eliminated papers that did not report empirical evidence of evaluating a behavioral intervention, reported in SDT terms. That left seven studies in four articles.

Figure 5-1 contrasts d’ and c for these studies, before and after the intervention. In this small sample of studies, the interventions were more effective at improving d’ for phishing detection (black circles), compared to the other contexts (blue squares), while having similar effects on c. For improving d’, the most effective intervention was embedded training. For decreasing c, a burst of high base rate training with feedback was most effective. Few studies reported individual variation in intervention effectiveness. However, given the heterogeneity of baseline performance (Chapter 3), it seems plausible that interventions might not influence all users equally. Our risk analysis allows for this possibility.

Table 5-1. Effectiveness of interventions in the literature.

Reference	Intervention	Task	Δ d'	Δ c
Kumaraguru et al. (2010)	Educational materials	Phishing detection	0.62	-0.54
Kumaraguru et al. (2010)	Embedded training (PhishGuru)	Phishing detection	1.73	-0.52
Kumaraguru et al. (2010)	Game in lab (Anti-Phishing Phil)	Phishing detection	1.09	0.00
Kumaraguru et al. (2010)	Game in field (Anti-Phishing Phil)	Phishing detection	0.97	0.37
Ben-Asher & Gonzalez (2015)[a]	Expertise	Network attacks	0.07	0.06
Wolfe et al. (2007)[a]	Burst of high base rate with feedback	Baggage screening	-0.49	-0.95
Bisseret (1981)[b]	Experience	Air traffic control	0.02	-0.18
		Average Effect Size	*0.57*	*-0.25*

[a] Reported hit and false alarm rates, converted to d' and c
[b] Reported as β and converted to c where $c = \ln(\beta)\ /d'$
Note: See Stanislaw & Todorov (1999) for more details on calculation of SDT parameters.

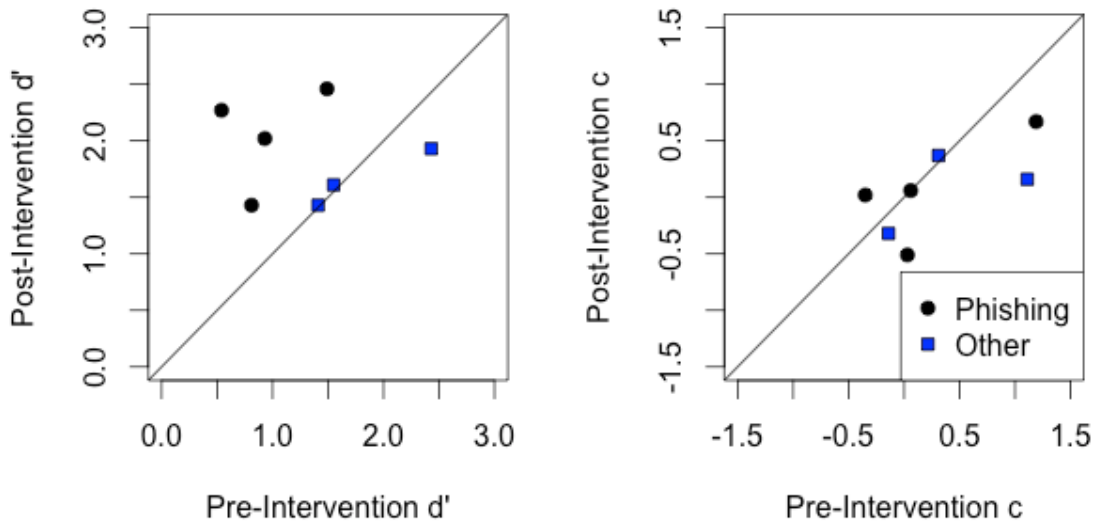

Figure 5-1. Average change in (a) d' and (b) c for various behavioral interventions.

Although the vulnerability of a system is determined by its users' d' and c, system operators may be concerned about the realized number of successful attacks. That rate will partially determine the total cost to their system from such

attacks and the appropriate investment in their reductions. We assess performance in terms of the number of successful phishing attacks (out of 100). Figure 5-2 shows performance for different values of d' and c. When c is negative and d' is high, the risk is low (blue). When c is positive and d' is low, the risk is high (red). As seen in the figure, users can have the same number of successful attacks with varying SDT parameters. For example, a user with d' = 1.25 and c = 0.31 has the same number of successful attacks as a user with d' = 0 and c = -0.32.

The black circles in Figure 5-2 show the vulnerability associated with each individual participant in Chapters 3 and 4, as determined by their d' and c values (for the behavior task, which more closely captures the actions affecting system performance, than does the detection task). The risk model in the next section assesses the value of behavioral interventions for users at different vulnerability levels (analogous to the color bands in Figure 5-2). It defines benefits in terms of reduced vulnerability (i.e., a lower rate of successful attacks) and costs in terms of those associated with any increased rate of false alarms, which reduce users' ability to do their jobs (and might reduce the effectiveness of the intervention over time), as well as the costs of services for implementing behavioral interventions.

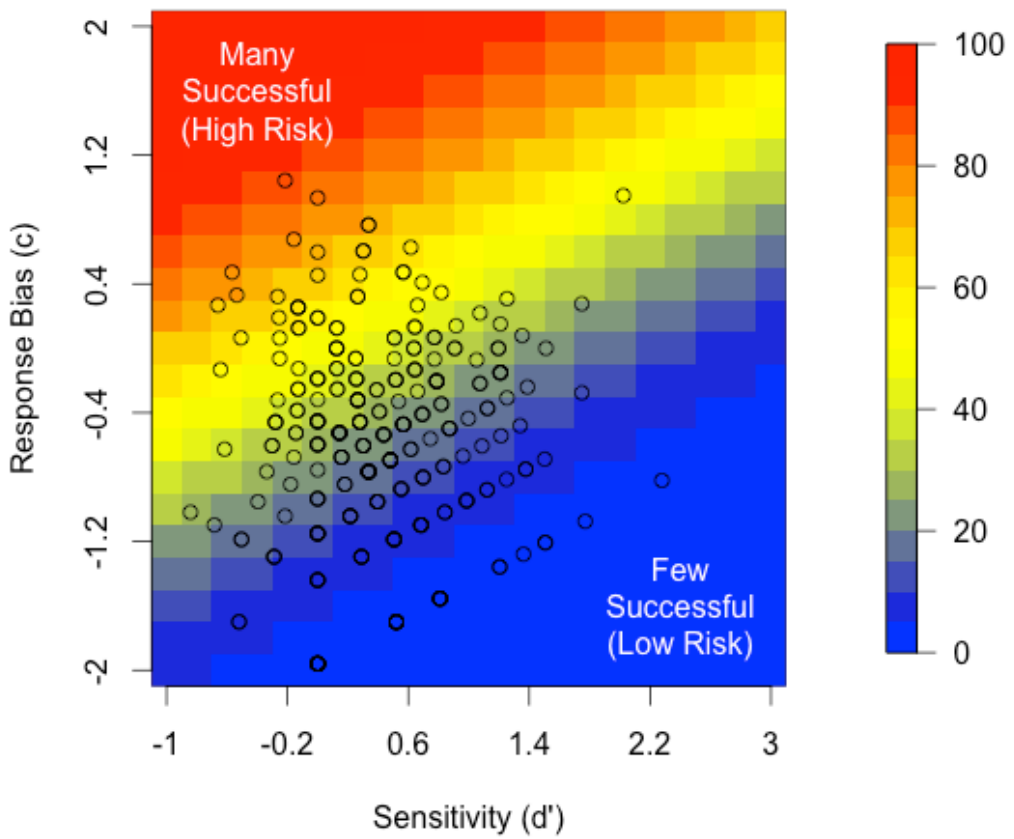

Figure 5-2. Number of successful phish out of 100 (denoted by color) as a function of d' and c. Observations from Chapters 3 and 4 are plotted in black. Risk is high when d' is low and users are biased toward clicking on links in emails (positive c).

The model recognizes the natural variation in phishing susceptibility, as well as the fact that users can have the same level of vulnerability for different reasons (i.e., combinations of d' and c), as seen in Figure 5-2. Interventions have different effects on the two SDT parameters (Table 5-1). As a result, they can have different effects on the vulnerability of users with the same performance. We use a simulation to (1) describe poor detectors, defined as the bottom 10% of users; (2) determine the cumulative contribution of those poor detectors to overall system vulnerability, and (3) compare the benefit-cost of behavioral interventions when focused on poor detectors or all users.

5.2. Method

5.2.1. Overview of Risk Simulation

The present model simulates the effect of behavioral interventions on users' phishing susceptibility for two different types of attacks: random (with no special recognition of the target) and spear phishing (with some personal information). As depicted in Figure 5-3, in each iteration of the model, we first generate a sample of individuals with varying vulnerability, defined by d' and c, drawn from the distribution of empirical estimates in Chapter 3 & 4 (Step A). We then estimate each user's initial (or baseline) performance, in terms of the number of phishing emails that they fall for (misses) and the number of legitimate emails that they mistake for phishing (false alarms) (Step B). For each user, we sample an intervention from a Normal distribution defined by the literature review (of both phishing and non-phishing interventions) in Table 5-1 (Step C). That distribution is used to reflect the variability in the effects of these interventions on individual users (which is not routinely reported in studies). We then recalculate that user's d' and c, incorporating the intervention's effects (Step D).

In Steps B and D, we estimate vulnerability separately for two types of attacks, random and spear phishing. The ability to detect random phishing attacks is determined by the users' initial d' and c as well as the effects of any intervention. Because spear phishing emails are specifically designed to look like legitimate emails, users have a lower sensitivity (d'). The extent of that reduction in d' depends on how well the spear phishing email is crafted. As a placeholder for empirical estimates, the model uses a difficulty factor, f, ranging from 0, for a spear phishing

attack that is impossible to detect, to 1, for one that is no more difficult than a random phishing attack to detect. In the simulations reported here, the value for f is sampled from a uniform distribution over [0,1].

We assess performance on each simulated email as a draw from a Bernoulli distribution with some probability, where P_M is the probability of falling for a phishing email and P_{FA} is the probability of mistaking a legitimate email for phishing. This procedure is repeated for each email, both phishing and legitimate, that a user receives. In the simplest scenario (i.e., no interventions or spear phishing attacks), P_M is a function of initial vulnerability (d' and c):

$$P_M = 1 - \Phi(0.5d' - c)$$

where Φ represents a standard Normal distribution that converts a z-score to a probability (Macmillan & Creelman, 2004). In a scenario with an intervention having estimated impacts $\Delta_{d'}$ and Δ_c, and a spear phishing difficulty factor f, P_M is:

$$P_M = 1 - \Phi\{0.5[(d' + \Delta_{d'})f] - (c + \Delta_c)\}$$

These variables are summarized in Table 5-2, along with the basis of the parameter values used in the simulation. We report users' vulnerability (i.e., their probability of falling for a phishing attack, P_M) by decile to faciliate comparison between low and high performing users. Users in a low decile have a high probability of falling for attacks, while users in a high decile have a low probability (in effect, going down and to the right in Figure 5-2). We assume a 1% base rate, so for every phishing email, there are 99 legitimate emails. We estimate false alarms per user as P_{FA}:

$$P_{FA} = \Phi\{-0.5[d' + \Delta_{d'}] - (c + \Delta_c)\}$$

We report performance (or phishing accuracy) in terms of the rate of phishing emails that are missed. High-performing users fall for many attacks, while low-performing users fall for few. We make a distinction between expected vulnerability (estimated probability based on d' and c) and observed performance (simulated as draws from a Bernoulli distribution) in order to emphasize that, in this procedure, observations may not perfectly reflect reality (as defined by d' and c).

We use a Monte-Carlo simulation to incorporate uncertainty by assigning a distribution to each parameter. The results represent the outcome of 1,000 iterations, each involving 100 phishing attacks against 100 users with a 1% base rate. We compare different scenarios in terms of their benefit-cost (or net benefit), as described in the following section.

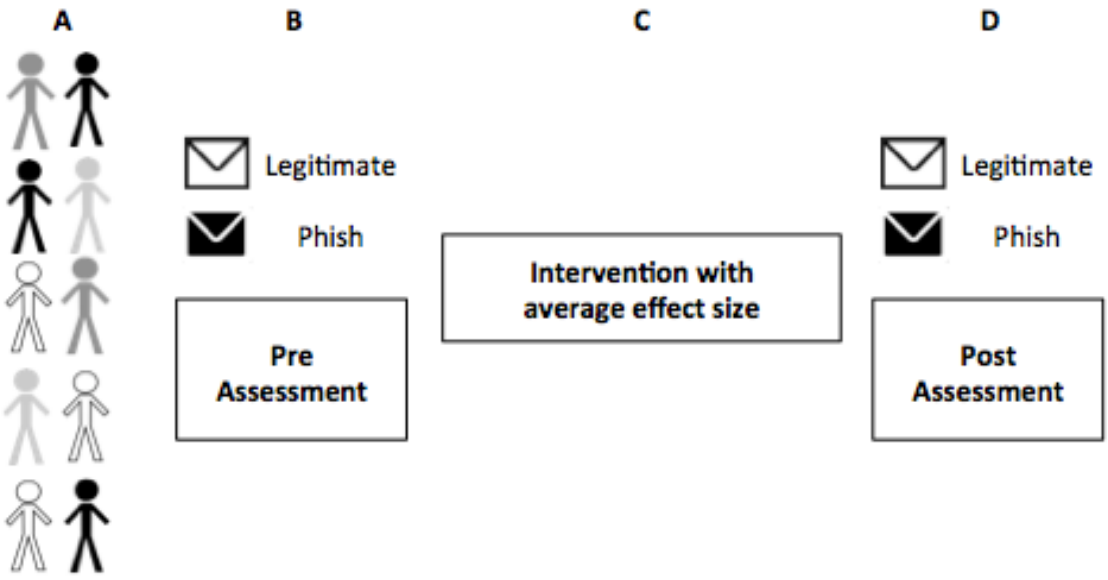

Figure 5-3. High level diagram of model.

Table 5-2. Model Inputs.

Inputs	Value	Description
Difficulty factor	$f \sim U(0,1)$	For random phishing attacks, f = 1. For spear phishing attacks, f ranges from f = 0, which eliminates d', to f = 1, which preserves it.
Sensitivity	$d' \sim N(0.4, 0.5)$	Estimated from experimental data (Chapters 3 & 4)
Response bias	$c \sim N(-0.6, 0.65)$	Estimated from experimental data (Chapters 3 & 4)
Effect on d'	$\Delta_{d'} \sim N(0.57, 0.76)$	The mean and standard deviation was determined based on the literature review (see Table 5-1).
Effect on c	$\Delta_{c} \sim N(-0.25, 0.45)$	The mean and standard deviation was determined based on the literature review (see Table 5-1).

5.2.2. Benefit-Cost Analysis

Benefit-cost analysis is a systematic, analytical approach for assessing trade-offs between options (Lave, 1996). We use this technique to estimate the difference between the benefits and costs of anti-phishing interventions. We account for both the direct cost of the intervention (e.g. usage fees, lost time and productivity) as well as any indirect costs arising from behavior change (e.g. increased false alarms). In this case, behavior change is measured as the difference in the number of successful attacks and false alarms. When that change is negative, an intervention reduced successful attacks and false alarms, giving it net benefits.

The cost of successful attacks could be as low as the cost of changing a compromised password or as high as a high profile data breach. The cost of a false alarm could be as low as typing a URL into a browser (rather than clicking on the link) or as high as a lost business opportunity. We assume that probabilities are not uniform across the range of possible impacts, but that high cost events are rare.

Therefore, we model the costs of attacks and false alarms with a lognormal distribution, which has a long positive tail to accommodate those rare, high-cost events. Table 5-3 summarizes these assumptions. Appendix C presents descriptive statistics for each input.

Table 5-3. Summary of assumptions for benefit-cost analysis.

	Cost	**Benefit**	**Value**	**Source**
Attack	Additional successful attacks	Avoided attacks	1,800* LogNormal (0,1)	Cyveillance (2015)
False Alarm	Additional false alarms	Avoided false alarms	LogNormal (0,2)	
Intervention	Cost of implementation; Lost productivity	N/A	U(1,10); U(10,100)	Ponemon (2016); Range accounts for time spent on intervention (1-60 minutes), frequency (1-52 times/year), and hourly wage for professionals ($20-50).

5.3. Results & Discussion

The results are presented in three sections. Section 5.3.1 compares observed performance and expected vulnerability. Section 5.3.2 assesses the cumulative vulnerability by decile of users' vulnerability. Section 5.3.3 examines the costs and benefits of behavioral interventions for different deciles of users.

5.3.1. Measurement of Vulnerability

This section compares observed performance to expected phishing vulnerability, (P_M), in order to identify the characteristics of poor detectors. We

define a distribution of users, characterized by their relative proficiency as detectors (in deciles), as potential targets of selective interventions. We use the values of sensitivity (d') and response bias (c) observed in Chapter 3 and 4, taking the behavior task (rather than the detection task) because it is closer to users' actual tasks.

Figure 5-4a shows the performance of users (in terms of phishing accuracy, defined as the percent of phishing emails avoided) at each vulnerability decile. The means are monotonically related, by definition. Their slope proves to be relatively linear for the highest deciles (with the fewest successful attacks). However, they spread out for the lowest deciles, suggesting the potential value of targeting the poorest detectors. The estimates in Figure 5-4a assume the precision that comes with a test using 100 phishing emails. Tests with fewer data points will be more weakly related to vulnerability. Figure 5-4b and Figure 5-4c show the deciles of vulnerability as a function of each of the two SDT parameters separately. As would be expected, users in the 10th decile have relatively high d' and negative c, while users in the 1st decile have a low d' and positive c. For each decile, there is a wider range for d' than c, as reflected in a stronger correlation between P_M and c, $r(98) = 0.91$, $p < .001$, than with d', $r(98) = -0.38$, $p < .001$. This suggests that c is a more influential parameter than d' for interventions (across all users).

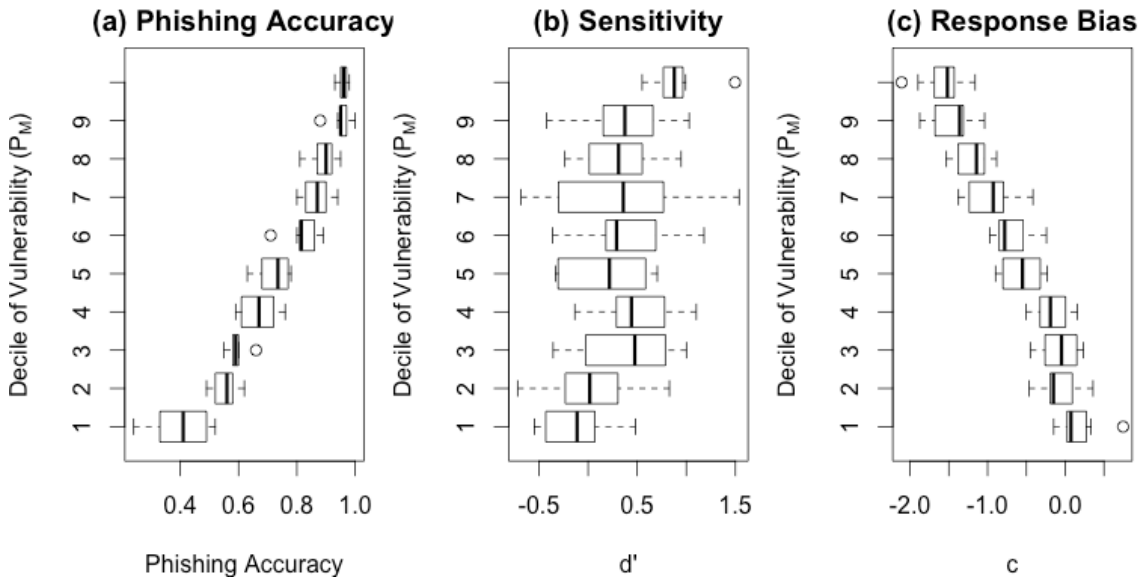

Figure 5-4. Decile of probability of falling for an attack as a function of (a) performance (accuracy) for 100 phishing emails, (b) sensitivity and (c) response bias.

5.3.2. Cumulative Vulnerability by Decile

Second, we assess cumulative vulnerability by decile, as a basis for evaluating the potential benefit of prioritizing poor detectors for behavioral interventions. Figure 5-5a translates the estimates of Figure 5-4a into cumulative distribution curves for the percentage of successful attacks. The black circles show these estimates for random phishing attacks, where the bottom 10% of users account for 26% of the total number of successful attacks. The blue triangles are for spear phishing attacks, which reduce d', where the bottom 10% of users account for 24% of successful attacks. The two curves are similar, despite the greater difficulty of distinguishing spear phishing attacks, because of the relatively weak relationship between d' and vulnerability (Figure 5-4b). Figure 5-5b shows the necessarily similar pattern for the number of successful attacks.

Figure 5-5b shows the number of successful attacks by decile. The total number of successful attacks is higher for spear than for random phishing attacks. This is particularly noticeable for high decile users, who fall for few random attacks. In contrast, low decile users fall for both types of attacks equally. Thus, spear phishing may be of particular concern for interventions targeting high decile users. In summary, poor detectors (bottom 10%) account for a disproportionate share of an organization's overall vulnerability for both random and spear phishing. This suggests that it may be worthwhile to focus intervention resources on poor detectors. This question is addressed in Section 5.3.3.

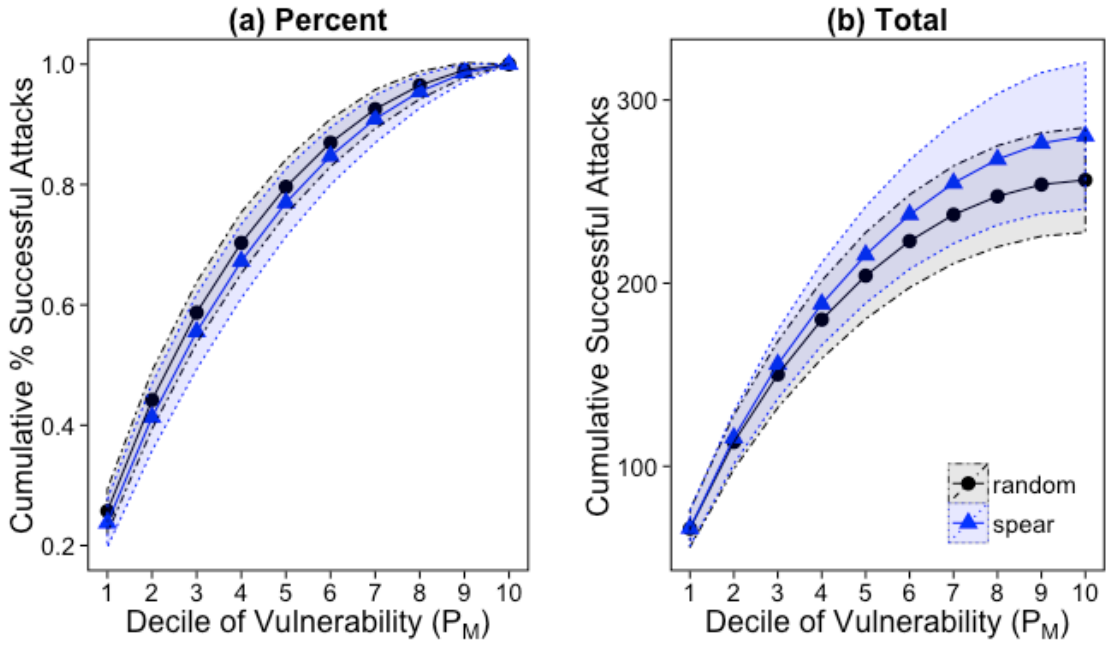

Figure 5-5. Cumulative (a) percent and (b) total number of successful attacks (i.e. performance) per vulnerability decile for 1 attack on 1,000 users. The error bars are +/- 2 standard deviations.

5.3.3. Benefit-Cost Analysis of Behavioral Interventions

Third, we evaluate the benefit-cost (or net benefit) of behavioral interventions. The benefits of an intervention are determined by the net reductions

in the numbers of successful attacks and of legitimate emails mistaken as phishing (false alarms). The costs of an intervention include those associated with its implementation (e.g. fees, lost productivity) and any additional successful attacks and false alarms that it unintentionally creates (e.g. by increasing trust in spam filters, which do not completely protect users). We first report results from a Monte-Carlo simulation varying the type of attack on the net benefit of an intervention when administered to users in each decile. We then report a sensitivity analysis examining the influence of our assumptions. As before, estimates of baseline performance are from the behavior group in Chapters 3 and 4. Estimates of intervention effects are taken from Table 5-2. Estimates of costs are from Table 5-3.

Figure 5-6 shows the net benefit of interventions, using these estimates, when applied to users in each decile, for random and spear phishing attacks. Given the fixed costs of the intervention (per user), the net benefits are much greater for users in the lower deciles, who contribute a disproportionate share of the system's vulnerability (Figure 5-5). However, some benefit exists even with the best detectors. For low decile users, the net benefit is somewhat greater for random attacks because they are easier to detect, so that interventions have a larger effect. Because low decile users fall for more attacks overall, the difference is larger. For high decile users, the net benefit is slightly greater for spear phishing attacks because they are more likely to be successful than a random attack (Figure 5-5b). Therefore, any benefit from an intervention is more likely to be realized. For random attacks, the mean net benefit is $580,000 (SD = $220,000) for users in the 1st performance decile, or 20% of the total net benefit of a system-wide program. It

is \$56,000 (SD = \$50,000) for users in the 10th decile, or just 2% of the total net benefit. For spear phishing attacks, the mean net benefit is \$440,000 (SD = \$180,000) for users in the 1st performance decile, or 18% of the total benefit. For users in the 10th decilie, the mean net benefit is \$60,000 (SD = \$48,000) or 2% of the total net benefit. Ultimately, the net benefit is positive (above the dotted line in Figure 5-6), under most conditions for all users. The next section investigates the effects of sensitivity analyses, varying model parameters.

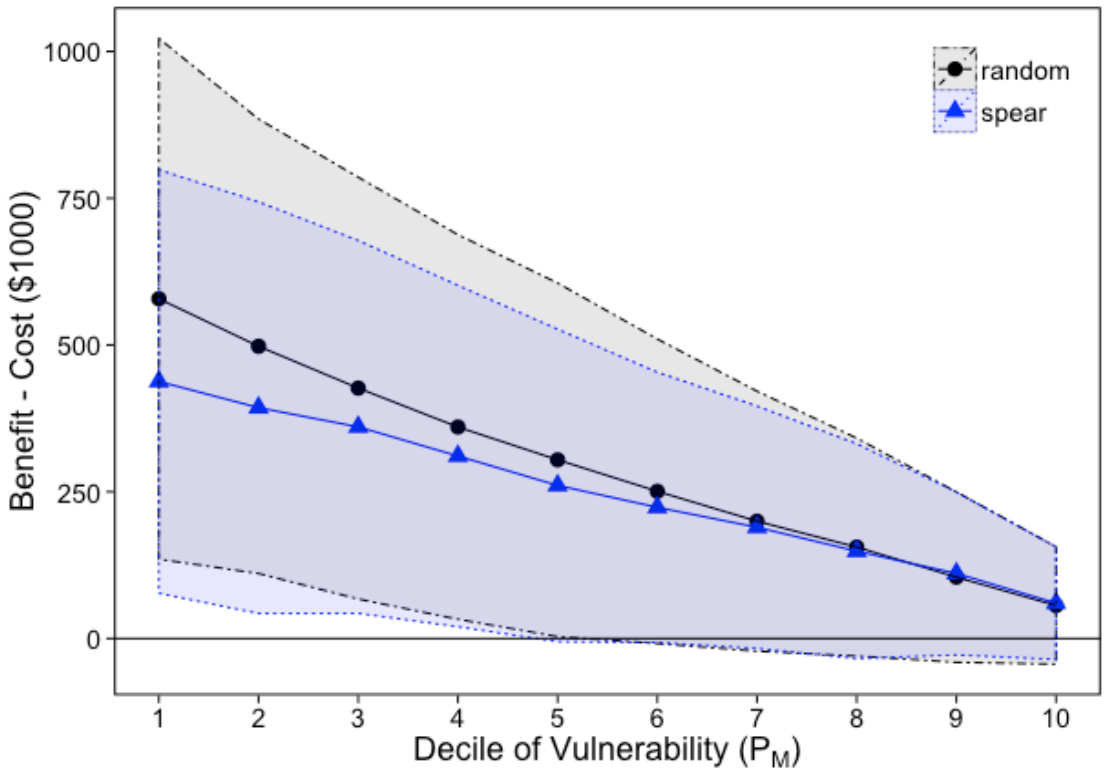

Figure 5-6. Benefit-cost per decile of performance where scenarios above the 0 line have positive net benefit.

5.3.3.1. Sensitivity Analysis

Table 5-4 shows the sensitivity of our estimates to varying each parameter in the model independently, for users in each decile, as a percent change from baseline

performance (without the intervention). Each row represents a parameter that was varied. The baseline assumptions are the mean inputs from the Monte Carlo model. The worst and best scenario assumptions are either the minimum or maximum inputs from the Monte Carlo model or other values of interest as noted in the text below. The first column in Table 5-4 provides the baseline assumptions, yielding a net benefit of $610,000 for the 1st decile and $63,000 for the 10th decile. Since the results are reported in terms of percent change of net benefit from the baseline, cases where is is less than -100% (bolded in Table 5-4) are where the benefit-cost crosses 0.

Row 1 shows the effects of varying the mean sensitivity (d') of users across the nominal range of d' (0 to 3). The baseline scenario used 0.4, the mean d' observed in the behavior task for Chapters 3 and 4. When the initial mean d' of users was very poor (d' = 0), high decile users benefited from an intervention more than low decile users because they could be more responsive (since they already had some, rather than no, ability to detect phishing emails. Even with the strongest of interventions, low decile users' d' still reflected weak discrimination. When the initial mean d' of users was very high (d' = 3), interventions had limited net benefit for high decile users, who were already able to almost perfectly distinguish between phishing and legitimate emails.

Row 2 varies the mean response bias (c) of users across the nominal range of c (-2 to 2). The baseline scenario used -0.6, the mean c observed in the behavior task for Chapters 3 and 4. At the worst-case value, where users were very incautious (c = 2), the benefits were much greater for high decile users. Low decile users were so

incautious to begin with (at the baseline value) that the intervention still left them falling for many attacks. At the best-case value, where users were already very cautious (c = -2), the intervention had little net benefit for all users.

Rows 3 and 4 show the results of sensitivity analyses varying the effectiveness of interventions. The worst-case values were chosen to represent interventions that not only failed, but backfired, significantly reducing d' or increasing c ($\Delta_{d'}$ = -1, Δ_c = 1). They had net costs (rather than benefits). The best-case values were chosen to represent very effective interventions, decreasing or increasing d' and c by 1. They led to increased net benefits over baseline. Interventions that increased d' ($\Delta_{d'}$ = 1) provided greater net benefit for all users, compared to the baseline. Interventions that decreased c (Δ_c = -1) increased net benefit for low decile users, but decreased net benefit for high decile users (due to increased false alarms). Thus, interventions must be careful to not induce unwanted behavior by reducing detection performance (by decreasing d' or increasing c) or increasing false alarms (by decreasing c too much for high decile users).

The final three rows vary the financial costs of successful attacks, false alarms and interventions. We assessed the worst and best values used in the Monte Carlo model (except for the worst intervention cost, a $10,000 intervention was not used in the Monte Carlo model, but represents the minimum cost for the net benefit to be negative). Very expensive attacks (e.g. costing $200,000/user affected) and false alarms (e.g. costing $100,000 per email) make any behavioral interventions seem extremely cost-effective since interventions generally reduce successful attacks and false alarms. However, these types of events are unusual, so they

provide little guidance. If there is no cost of an attack, there is little net benefit, except from avoided false alarms. If there is no cost of a false alarm, the net benefit does not change from the baseline, suggesting that the vast majority of the estimated net benefit can be attributed to avoided attacks. The cost of the intervention would outweigh the benefits for high decile users when it approaches $10,000 per person. These results are summarized in Table 5-4.

Table 5-4. Percent change of mean benefit-cost for random attacks from the baseline scenario (reported for the 1st and 10th decile).

Parameter	**Baseline**	**Worst**	**1st B - C**	**10th B - C**	**Best**	**1st B - C**	**10th B - C**
1. Mean d'	0.4	0	-4%	43%	3	-46%	-85%
2. Mean c	-0.6	2	-97%	820%	-2	-51%	-98%
3. Effect on d'	0.57	-1	**-170%**	**-290%**	1	46%	64%
4. Effect on c	-0.25	1	**-170%**	-73%	-1	92%	-75%
5. Cost of Attack	$3,000	$0	-99%	-94%	$200,000	6,500%	6,100%
6. Cost of FA	$7	$0	-1%	-9%	$100,000	18,000%	100,000%
7. Cost of Intervention	$60	$10,000	-16%	**-160%**	$10	-1%	0%
1st B - C	$610,000						
10th B - C	$63,000						

Note: Each parameter is varied independently while the other parameters are held at the baseline value. Cases where the change in net benefit is less than -100% (bolded) are where the benefit-cost crosses 0.

5.4. Conclusion

In this study, we used a Monte Carlo model to assess the value of implementing anti-phishing behavioral interventions under a wide range of scenarios. First, we identified poor detectors, defined here as the bottom 10% (or 1st decile). Second, we assessed the cumulative vulnerability due to poor detectors. Lastly, we performed benefit-cost analyses and assessed the sensitivity our our estimates to modeling assumptions. Overall, this work suggests that it is beneficial to (re)allocate resources to poor detectors to reduce their susceptibility. An

organization can do this by targeting poor detectors (if measurement is sufficiently precise to identify them), using interventions that disproportionately improve the performance of poor detectors, or using interventions that affect all users equally (accepting the risk that high decile users may have more false alarms). Our three primary findings are:

First, poor detectors tend to have both low d' and high c (indicating that they treat most emails as legitimate). Between those two parameters, c is much more closely related to vulnerability, suggesting that interventions should focus on c. In order to assess performance, system operators must collect data on how users treat phishing emails. Due to the random nature of susceptibility to phishing emails, it is difficult to distinguish between users who are bad at detecting phishing emails because they have high vulnerability or seem bad at detecting phishing emails because they are unlucky without sufficient data points. The more data they collect, the closer performance is to true vulnerability.

Second, by definition, poor detectors create a disproportionate amount of the overall risk. The simulation estimates just how great that share is. Under the model assumptions, it is quite large, suggesting the potential value of targeting them. Spear phishing attacks are more difficult for all users to detect. However, the difference is only observable for high decile users since they fall for so few attacks. Low decile users fall for random and spear phishing attacks equally. This suggests that high decile users may benefit from interventions specifically related to spear phishing.

Third, the benefit-cost analysis suggests the value of focusing resources on the more susceptible users – although under the baseline conditions, there is net

benefit for almost all users. The parameter estimates here were drawn from direct observation (Chapter 3 and 4), the research literature, or assumptions about normal organizational conditions. For the conditions that they describe, the net benefit is much higher for low decile than high decile users. Interventions may have a negative net benefit if they increase false alarms (beyond the benefit of avoided attacks) or inadvertently decrease d' or increase c. For example, a spam filter might increase c if users believe that they don't need to watch out for phishing emails because the spam filter will catch them.

This research has several limitations. One is that the costs of behavioral interventions are not fully modeled. For example, we do not account for such qualitative costs as annoying employees. In a study of phishing using social connections, Jagatic et al. (2007) faced a strong backlash for using people's real names as the senders of the fake phishing emails. An organization could face similar issues when attempting to train users to detect such attacks. Caputo et al. (2014) report that some of their users felt ashamed about clicking on the embedded training. Users may also be annoyed if the intervention is time-consuming or boring (Herley, 2014). Similarly, there may be benefits, such as increased reporting of phishing emails, that we do not quantify. Also, as noted, the behavioral estimates were from performance on an experimental task, rather than actual experience.

Future work could explore alternative ways of modeling spear phishing. Here, we model it as a reduction in d' based on the vigilance concept of similarity (Lynn & Barrett, 2014). Kaivanto (2014) modeled spear phishing using a parameter called "match quality," a binary factor that indicated a reduced d' rather than a

random reduction, as modeled here. However, it is possible that, in some settings, spear phishing messages may influence c, for example, by creating a sense of urgency or tapping into human emotions, such as greed (Vishwanath et al., 2011). This may motivate a user to perceive the email as legitimate even if they wouldn't normally.

Overall, these analyses suggest that the net benefit may be maximized when behavioral interventions approach users differently based on their vulnerability. Low decile users may benefit most from interventions designed to reduce their response bias (c). High decile users are more at risk for spear phishing because it undermines their otherwise low susceptibility. The same intervention may not help both of these types of users. There is already interest in the security community in tailoring behavioral interventions to improve security (Egelman & Peer, 2015). However, one of the main challenges is predicting which kind of a user a particular individual is.

A second challenge, which was beyond the scope of this study, is determining which interventions are effective for which users. At present, most studies report the average improved performance for their intervention. However, it may be useful to report the average improvement per decile to better understand the interaction between changes in d' and c. It may be more cost-effective to use behavioral interventions that have the largest effect size for low decile users, even if there is a cost to giving the intervention to high-decile users who won't benefit much.

6. Conclusions

6.1. Approach

This thesis bridges the vigilance and computer security literatures, in order to improve the measurement, management, and understanding of phishing susceptibility. I extended previous research (Kumaraguru et al., 2010) applying signal detection theory (SDT) to phishing susceptibility with a combination of experiments, correlational analysis, and simulations.

In Chapter 2, I summarized relevant research on vigilance and phishing. Vigilance is the ability to remain alert in order to detect small changes or rare stimuli over time (Mackworth, 1948). Researchers have found that vigilance performance can be influenced by task, environmental and individual factors (Ballard, 1996), which roughly parallels the factors identified in the phishing susceptibility literature. Although many researchers have studied individual and environmental factors that influence phishing susceptibility (Pattinson et al., 2012; Sheng et al., 2010; Vishwanath et al., 2011; Welk et al., 2015; Wright & Marett, 2010), less work has been done on task factors. This work aims to address this gap and show the value of framing phishing detection as a vigilance task.

In Chapter 3, I measured phishing susceptibility for two interrelated tasks, detection ("is this a phishing email?") and behavior ("what would you do if you received this email?"), in an online experiment. I manipulated three task variables: (1) which task comes first, detection or behavior (Experiment 1); (2) whether participants perform both tasks (Experiment 1) or just one (Experiment 2) and (3) whether participants are told, or must infer, the base rate of phishing messages.

In Chapter 4, I assessed the validity of the experimental measurement from Chapter 3. Using participants and data from the Security Behavior Observatory (SBO), an existing effort to measure computer users' security habits over time (Forget et al., 2014), I evaluated (1) face validity by repeating the experimental tasks from Chapter 3 with a community sample, (2) construct validity by correlating SDT parameter estimates with scores on the Security Behavior Intentions Scale (SeBIS) (Egelman & Peer, 2015), and (3) predictive validity by comparing experimental performance to adverse outcomes experienced by users on their home systems, namely, visits to malicious websites and presence of malicious files.

In Chapter 5, I used a risk-analytic simulation to estimate the value of behavioral interventions for users of varying vulnerability. These analyses (1) identified which users were most vulnerable, (2) assessed the relative risk attributed to the most vulnerable users, (3) estimated the benefit-cost of behavioral interventions by vulnerability level, and (4) evaluated sensitivity to random versus spear phishing.

6.2. Findings

In Chapter 3, I found that users employed much more cautious decision-making strategies for the behavior task than for the detection task. On average, users had a lower d' and c for the behavior task than for the detection task. However, users still fell for phishing attacks, suggesting they did not sufficiently compensate for their limited detection ability, despite showing some sensitivity to the extent of that ability (as expressed in their confidence judgments) and their assessment of the consequences of misses. Individual performance varied widely,

but group-level performance on the laboratory experiment was robust across experimental conditions. There was no effect for changing the order of the tasks, whether participants performed one task or both, or whether they received notification of the base rate. These patterns suggest the generalizability of results across laboratory settings. In addition it may be possible to design interventions focused on users' confidence and perceptions of consequences.

In Chapter 4, there was evidence to support the validity of the behavior task, but the results were largely inconclusive. The experimental findings from Chapter 3 generalized to the community population and exhibited the same variance in individual ability. The tendency to not click on links (negative behavior c) was correlated with a validated scale of security behavior intentions (SeBIS), providing some evidence of construct validity. This suggests that participants who reported looking at the URL before clicking on a link (SeBIS) tended not to click on links in emails (behavior c). There was no evidence of predictive validity, in that there was no relationship between performance on the experiment and real world outcomes, including visits to malicious URLs and presence of malicious files. I provided recommendations for validating scales using this type of novel real-world data.

In Chapter 5, I showed the value of assessing phishing susceptibility in terms of SDT to design better behavioral interventions. I found that the most susceptible users had a positive c and low (although variable) d'. In addition, they represented a disproportionate amount of overall phishing risk for both random and spear phishing attacks. In a Monte-Carlo model, benefit-cost analysis indicated that the net benefit of behavioral interventions was much higher for more susceptible users.

However, the net benefit of interventions for the least susceptible users was still positive under most conditions. Overall, this work suggested that the system-level net benefit may be maximized when behavioral interventions approach users differently based on their true vulnerability.

6.3. Scientific Contributions

In summary, this thesis shows the applicability of vigilance research as a framework for understanding phishing susceptibility and evaluating anti-phishing behavioral interventions, while extending that literature in this distinctive domain – which involves an intellectually demanding task, done concurrently with users' main task (unlike, say, baggage screening, where detecting deception is the primary task). Vigilance research has identified task, environmental and individual factors that can affect detection ability (Ballard, 1996). This thesis primarily contributed to understanding of task factors. In Chapter 3, I evaluated the nature of the task (detection versus behavior), payoffs (or perceived consequences), and information about the base rate. Within subjects, users used much more cautious decision-making strategies (lower d' and c) for the behavior task, where the consequences of falling for an attack are much more salient, in order to reduce misses at the cost of increasing false alarms. Between subjects, users who perceived worse consequences of falling for an attack were also more cautious (as measured by a lower c). Our experimental manipulation of base rate information had no effect, suggesting robustness across experimental designs. Our results were robust across different populations (community vs. mTurk populations assessed in Chapter 4) and

experimental manipulations (e.g. performing the detection and behavior tasks individually rather than together in Chapter 3).

Future research should refine a method of evaluating phishing susceptibility in terms of SDT. This may involve measuring susceptibility at a more realistic base rate. To reduce cost, it may be worthwhile to investigate how low the base rate of phishing emails needs to be to reflect true vulnerability. For example, a base rate of 20% may be sufficiently low to induce measurable changes in response bias (c).

In terms of improving anti-phishing interventions, future research could evaluate other ways to influence task factors, which have been effective at improving vigilance in other contexts (Wolfe et al., 2007). For example, embedded training uses partial feedback so users only reassess their decision strategy when they fall for a fake phishing email. It may be effective to send congratulatory messages to users who do not fall for the fake phishing email in order to maintain their low c. In addition, it may be illuminating to assess other interventions in terms of SDT. For example, warning messages that include sanitized phishing emails may increase awareness of the base rate of phishing (decrease c), but may also provide less interaction for the user (lower magnitude of effect) and may inappropriately increase trust in technical defenses, such as the spam filter (increase c).

6.4. Practical Contributions

I also showed how SDT parameters can be used to evaluate (and design) anti-phishing interventions, in the context of a quantitative risk analysis. I created a model that quantified the benefit-cost of changing users' behavior as a function of their initial vulnerability to answer the following question: how does the wide range

of user vulnerability interact with (a) the effectiveness of an intervention and (b) the type of attack?

I found that users of different vulnerabilities (as assessed by d' and c) were affected differently by interventions. In general, the net benefit was much higher for users who were more vulnerable. However, an intervention that helps more vulnerable users (e.g. by causing a large decrease in c) could reduce the already small net benefit for less vulnerable users by increasing their likelihood of perceiving legitimate emails as phishing (and reducing their productivity). In addition, I found that the benefit-cost of interventions for less vulnerable users increased slightly when it helped them avoid spear phishing attacks, which are harder to detect. This did not hold for more vulnerable users, who fell for fewer random phishing attacks after an intervention, but still struggled to avoid spear phishing attacks. Together, these results suggest that measuring phishing vulnerability in terms of signal detection theory paints a clearer picture of the interaction between user vulnerability, effectiveness of interventions, and types of threats.

However, measuring phishing susceptibility in terms of SDT is still a challenge. In Chapter 4, I was unable to conclusively validate the measurement method used in Chapter 3. The complexity of the real world, where technology and user ability interact to protect (or threaten) computer security, inhibited my ability to measure a relationship between performance in the laboratory study and real world outcomes (such as visits to malicious URLs and presence of malicious files). In this situation, it is possible a laboratory measurement may be more useful for

describing human vulnerability than what happens in the real world, which will often be confounded by other variables.

Future work could address this question by comparing measurement of vulnerability in an embedded (field) versus laboratory setting. Embedded measurement is limited by the frequency of phishing emails, so it may take a long time to collect enough data to assess vulnerability. However laboratory settings lack the ecological validity of a user's real inbox. Understanding the differences in how these methods bias measurement has implications for the evaluation of interventions and assessment of real-time vulnerability for risk analysis. Embedded measurement may be more useful for the former than the latter.

This work also has implications for the design of behavioral interventions. Chapters 4 and 5 suggest that interventions that decrease c (how suspicious an email must be to avoid clicking on the link or attachment) may be more effective than ones that increase d' (ability to discern the difference between phishing and legitimate emails). The results of Chapter 3 indicate that behavior c is negatively correlated with time spent on each email and perceived consequences, as well as being positively correlated with confidence. Future work should investigate the causal relationships between these variables to identify potential interventions.

In addition, this work suggests an alternate explanation for how embedded training works. Rather than expecting users to show continuous improvement, operators should expect all users to fall for the fake phishing emails occasionally. These failures need not indicate that their ability to identify phishing emails (or d') is deficient. Rather, it could reflect users' expected sensitivity to perceived changes

in base rate. If they don't notice any phishing emails for an extended period of time, they could rationally adjust their criteria (or c) for how suspicious an email must be to avoid clicking on the link. Fake phishing emails (via embedded training) may serve to readjust the perceived base rate as needed to maintain desired performance. If users never fall for fake phishing emails, that could mean that (a) the emails are too easy to detect, (b) the emails are not providing any additional information about the task (e.g., information about the base rate), or (c) the user has a very negative c and avoids clicking on most links and attachments.

6.5. Policy Implications

Industry and government stakeholders are broadly interested in measuring phishing susceptibility, implementing effective anti-phishing interventions, and assessing phishing risk. I identify three potential audiences for this research including (a) corporate security officers, (b) government regulators, and (c) anti-phishing companies.

Corporate security officers should employ anti-phishing interventions for all employees. Chapter 5 shows that under the assumptions of our risk model, the benefits of anti-phishing interventions exceed the costs for all users, regardless of vulnerability, under most conditions. This recommendation might no longer hold in cases where the probability of a high-consequence false alarm (ignoring a legitimate email because it is perceived as phishing) is sufficiently high to outweigh the probability of a high-consequence phishing attack. This can be out-sourced to anti-phishing companies or provided in-house. Interventions that improve knowledge of phishing cues (i.e. increase d') may not be sufficient to encourage cautious decision

strategies (i.e. lower c). Given the strong positive correlation between c and vulnerability observed in Chapter 5, it may be more important to reduce c than increase d'.

In many industries, such as critical infrastructure, cybersecurity practices are regulated to some extent. Often that regulation involves voluntary guidelines, rather than mandatory rules, such as the North American Electric Reliability Corporation's Critical Infrastructure Protection standards (NERC CIP). In both settings, there is increased interest in risk-based, rather than checklist, evaluations of compliance. Such analyses might require quantitative estimates of phishing susceptibility. SDT provides such estimates, making the essential distinction between users' abilities (d') and their decision rules (c). These two aspects of performance are conflated in measures that consider solely the number of successful attacks. How any intervention affects each aspect is an empirical question, which might be addressed by repeated assessment before and after an intervention (or over the course of time, as users acquire experience or adversaries change their tactics). How these estimated effects should be incorporated in system design is an analytical question, which I addressed in the risk assessment of Chapter 5.

I found only weak correlations between performance on the empirical test and real-world performance, as reflected on the Security Behavior Observatory (SBO), perhaps the most intense record of user behavior to date. As discussed in Chapter 4, it is unclear what those weak correlations say about the external validity of the experimental test for the specific vulnerabilities available in the rich SBO

record, given the difficulty of extracting a clear signal. These results pose a need for future research and a challenge for regulators to decide how to evaluate users, systems, and interventions.

Anti-phishing companies (or internal units) may choose (or be required) to describe their products (e.g. embedded training) in vigilance terms, in order to design interventions that best suit specific systems. For example, vigilance research suggests that changes in c last longer than changes in d' (Wolfe et al., 2013). Therefore, it may be more correct to describe embedded training as increasing the perceived base rate (which reduces c), rather than improving users' ability to detect phishing emails. In addition, there may be a market for personalized interventions, which might predict how often to send embedded training or vary the difficulty (e.g. random vs. spear phishing) of the fake phishing emails.

As phishing risk evolves, I anticipate a need for more tools to measure, manage, and understand phishing susceptibility. This thesis shows that vigilance, in terms of phishing detection as well as other contexts, provides particularly valuable tools.

7. References

Abrams, R., Pathak, J., Barrera, O., Ghimire, D. (2014). Browser Security Comparative Analysis: Socially Engineering Malware Blocking. NSS Labs. Retrieved from https://www.nsslabs.com/linkservid/A53E0FBA-5056-9046-9359282CBE961A0E/.

Appel, M. (2012). Are heavy users of computer games and social media more computer literate? *Computers & Education*, *59*(4), 1339–1349. http://doi.org/10.1016/j.compedu.2012.06.004

Apt, J., Lave, L. B., & Morgan, M. G. (2006). Power Play: A More Reliable U.S. Electric System. *Issues in Science and Technology*, 51–58.

Apt, J., Lave, L. B., Talukdar, S., Morgan, M. G., & Ilic, M. (2004). Electrical Blackouts: A Systemic Problem. *Issues in Science and Technology*, 55–62.

Ballard, J. C. (1996). Computerized assessment of sustained attention: A review of factors affecting vigilance performance. *Journal of Clinical and Experimental Neuropsychology*, *18*(6), 843–863. doi:10.1080/01688639608408307

Ben-Asher, N., & Gonzalez, C. (2015). Effects of cyber security knowledge on attack detection. *Computers in Human Behavior*, *48*, 51–61.

Birdwell, R. L., & Wolfe, J. M. (2013). If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening. *PLoS ONE*, *8*(5), e64366. http://doi.org/10.1371/journal.pone.0064366.s003

Bisseret, A. (1981). Application of signal detection theory to decision making in supervisory control: The effect of the operator's experience. *Ergonomics*, *24*(2), 81–94.

Boyce, M. W., Duma, K. M., Hettinger, L. J., Malone, T. B., Wilson, D. P., & Lockett-Reynolds, J. (2011). Human Performance in Cybersecurity: A Research Agenda. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 1115–1119. http://doi.org/10.1177/1071181311551233

Bravo-Lillo, C., Cranor, L. F., Downs, J. S., & Komanduri, S. (2011). Bridging the Gapin Computer Security Warnings. *IEEE Computer and Reliability Societies*, 18–26.

Butavicius, M., Parsons, K., Pattinson, M., & McCormac, A. (2015). Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails (pp. 1–11). Presented at the Australasian Conference on Information Systems, Adelaide, Australia.

Camp, L. J. (2009). Mental models of privacy and security. *IEEE Technology and Society Magazine*, *28*(3), 37–46. http://doi.org/10.1109/MTS.2009.934142

Caputo, D. D., Pfleeger, S. L., Freeman, J. D., & Johnson, M. E. (2014). Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy*, 28–38.

Carpenter, S., Zhu, F., & Kolimi, S. (2014). Reducing online identity disclosure using warnings. *Applied Ergonomics*, *45*(5), 1337–1342. http://doi.org/10.1016/j.apergo.2013.10.005

CERT, Insider Threat Team. (2013). *Unintentional Insider Threats: A Foundational Study* (CMU/SEI-2013-TN-022). http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=58744

Christin, N., Egelman, S., Vidas, T., & Grossklags, J. (2011). It's all about the Benjamins: An empirical study on incentivizing users to ignore security advice. In *International Conference on Financial Cryptography and Data Security* (pp. 16-30). Springer Berlin Heidelberg.

Coombs, C. H., Dawes, R. M. & Tversky, A. (1970) *Mathematical Psychology: An Elementary Introduction*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Cova, M., Kruegel, C., & Vigna, G. (2008). There is No Free Phish: An Analysis of "Free" and Live Phishing Kits. Presented at the USENIX Workshop on Offensive Technologies, San Jose, CA.

Cranor, L. F. (2008). A Framework for Reasoning About the Human in the Loop. *Usability, Psychology & Security*.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, 8(3), 1-18.

Cyveillance. (2015). *The Cost of Phishing: Understanding the True Cost Dynamics Behind Phishing Attacks*. Retrieved from http://info.cyveillance.com/rs/cyveillanceinc/images/CYV-WP-CostofPhishing.pdf.

Davinson, N., & Sillence, E. (2010). It won't happen to me: Promoting secure behaviour among internet users. *Computers in Human Behavior*, *26*(6), 1739–1747. http://doi.org/10.1016/j.chb.2010.06.023

DeCarlo, L. T. (1998). Signal Detection Theory and Generalized Linear Models. *Psychological Methods*, *3*(2), 186–205.

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, *54*(3), 304–313. doi:10.1016/j.jmp.2010.01.001

Dewitt, B., Fischhoff, B., Davis, A., & Broomell, S. B. (2015). Environmental risk perception from visual cues: the psychophysics of tornado risk perception. *Environmental Research Letters*, *10*(12), 1–15. http://doi.org/10.1088/1748-9326/10/12/124009

Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why Phishing Works (pp. 581–590). Proceedings of CHI, Montreal, Quebec, Canada.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are Your Participants Gaming the System? Screening Mechanical Turk Workers. Proceedings of *CHI*, Atlanta, Georgia.

Downs, J., Holbrook, M. B., & Cranor, L. F. (2006). Decision Strategies and Susceptibility to Phishing. Proceedings of Symposium on Usable Privacy and Security (SOUPS).

Drake, J., Mehta, P., Miller, C., Moyer, S., Smith, R. & Valasek, C. (2011). Browser Security Comparison: A Quantitative Approach. Accuvant Labs. Retrieved from http://files.accuvant.com/web/files/AccuvantBrowserSecCompar_FINAL.pdf.

Egelman, S., & Peer, E. (2015). Scaling the Security Wall (pp. 2873–2882). Presented at the 33rd Annual ACM Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/2702123.2702249

Egelman, S., & Peer, E. (2015). The Myth of the Average User (pp. 1–13). Presented at the NSPW, Twente, The Netherlands.

Egelman, S., Cranor, L. F., and Hong, J. (2008). You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In Proceedings of the CHI Conference on Human Factors in Computing Systems, (Florence, Italy), 1065–1074.

Egelman, S., Harbach, M., & Peer, E. (2016). Behavior Ever Follows Intention? In Proceedings of at CHI. http://doi.org/10.1145/2858036.2858265

Executive Order 13636. (2013). Improving critical infrastructure cybersecurity. Retrieved from https://federalregister.gov/a/2013-03915.

Farris, C., Treat, T. A., Viken, R. J., & McFall, R. M. (2008). Perceptual Mechanisms That Characterize Gender Differences in Decoding Women's Sexual Intent. *Psychological Science,* 19(4), 348–354.

Fischhoff, B. & MacGregor, D. (1986). Calibrating databases. *Journal of American Society for Information Sciences, 37*, 222-233.

Forget, A., Komanduri, S., Acquisti, A., Christin, N., Cranor, L., & Telang, R. (2014). Security Behavior Observatory: Infrastructure for Long-term Monitoring of Client Machines (CMU-CyLab-14-009). Pittsburgh: CyLab.

Franklin, J., Perrig, A., Paxson, V., & Savage, S. (2007). An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants. Presented at the Conference on Computer and Communications Security (CCS), Alexandria, VA.

Goodie, A. S., & Fantino, E. (1999). What Does and Does Not Alleviate Base-Rate Neglect Under Direct Experience. *Journal of Behavioral Decision Making*, *12*, 302–335.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. 1966. *New York*.

Hardee, J. B., Mayhorn, C. B., & West, R. (2006). I Downloaded What? : An Examination of Computer Security Decisions (pp. 1817–1820). Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51.

Herley, C. (2009). So Long, And No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. Presented at NSPW, Oxford, United Kingdom.

Herley, C. (2014). More Is Not the Answer. *IEEE Security & Privacy*, 14–19.

Hong, J. (2012). The state of phishing attacks. *Communications of the ACM*, *55*(1), 74. http://doi.org/10.1145/2063176.2063197

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd edition). John Wiley & Sons.

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A Comparison of Goodness-of-Fit Tests For the Logistic Regression Model. *Statistics in Medicine*, *16*, 965–980.

HTTP Archive. (2016) Trends. Retrieved on June 29, 2016 from http://httparchive.org/trends.php#bytesTotal&reqTotal.
Ion, I., Reeder, R., & Consolvo, S. (2015). "...no one can hack my mind": Comparing Expert and Non-Expert Security Practices (pp. 327–346). Presented at the Symposium on Usable Privacy and Security (SOUPS), Ottawa, Canada.
Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
Kaivanto, K. (2014). The Effect of Decentralized Behavioral Decision Making on System-Level Risk. *Risk Analysis*, *34*(12), 2121–2142. http://doi.org/10.1111/risa.12219
Kluger, A. N., & DeNisi, A. (1996). The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. *Psychological Bulletin*, *119*(2), 254–284.
Knoblauch, K., & Maloney, L. T. (2012). *Modeling Psychophysical Data in R*. New York, NY: Springer.
Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010). Teaching Johnny not to fall for phish. ACM Transactions on Internet Technology, 10(2), 1–31.
Lave, L. B. (1996). Benefit-Cost Analysis. In R. W. Hahn (Ed.), *Do the Benefits Exceed the Costs?* (pp. 104–134). New York: Oxford University Press.
Lichtenstein, S. & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*, 149-171.
Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.
Luo, X. R., Zhang, W., Burd, S., & Seazzu, A. (2013). Investigating phishing victimization with the Heuristic-Systemic Model: A theoretical framework and an exploration. *Computers & Security*, *38*(C), 28–38.
Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" Signal Detection Theory. *Psychological Science*. doi:10.1177/0956797614541991
Lynn, S. K., Wormwood, J. B., Barrett, L. F., & Quigley, K. S. (2015). Decision making from economic and signal detection perspectives: development of an integrated framework. *Frontiers in Psychology*, *6*. http://doi.org/10.3389/fpsyg.2015.00952
Lynn, S. K., Zhang, X., & Barrett, L. F. (2012). Affective state influences perception by affecting decision parameters underlying bias and sensitivity. *Emotion*, *12*(4), 726–736. http://doi.org/10.1037/a0026765
Mackie, R. R., Dennis, W. C. & Smith, M. J. (1994). Countering loss of vigilance in sonar watchstanding using signal injection and performance feedback. *Ergonomics*, *37*(7), 1157–1184. http://doi.org/10.1080/00140139408964895
Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6–21. doi:10.1080/17470214808416738
Macmillan, N. A., & Creelman, D. C. (2004). *Detection Theory: A User's Guide*. New York: Psychology Press.
Maddox, W. T. (2002). Toward a Unified Theory of Decision Criterion Learning in Perceptual Categorization. *Journal of the Experimental Analysis of Behavior*, *78*(3), 567–595.

Mason, W. & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods,* 44(1), 1-23.
Mayhorn, C. B., & Nyeste, P. G. (2012). Training users to counteract phishing. *Work*, *41*, 3549–3552. http://doi.org/10.3233/WOR-2012-1054-3549
Mickey, J., and Greenland, S. (1989). A study of the impact of confounder-selection criteria on effect estimation. American Journal of Epidemiology, 129, 125–137.
Miguel, E., Camerer, C., Casey, K. , Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R, Simmons, J. P., Simonsohn, U., and Van Der Laan, M. (2014). Promoting transparency in social science research. Science, 343, 30-31.
Mohan, D., Rosengart, M. R., Farris, C., Fischhoff, B., & Angus, D. C. (2012). Sources of non-compliance with clinical practice guidelines in trauma triage: a decision science study. *Implementation Science,* 7(103), 1–10.
Moore, T., & Clayton, R. (2007). An Empirical Analysis of the Current State of Phishing Attack and Defence. In Workshop on the Economics of Information Security (WEIS).
Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence*. Psychological Review*, *115*(2), 502-517.
Mumpower, J. L., & McClelland, G. H. (2014). A signal detection theory analysis of racial and ethnic disproportionality in the referral and substantiation processes of the U.S. child welfare services system. *Judgment and Decision Making*, *9*(2), 114–128.
Navalpakkam, V., Koch, C., & Perona, P. (2009). Homo economicus in visual search. *Journal of Vision*, *9*(1), 31–31.
Newell, B. R., & Rakow, T. (2007). The role of experience in decisions from description. *Psychonomic Bulletin & Review*, *14*(6), 1133–1139.
NIST. (2012). *Guide for Conducting Risk Assessments*. Retrieved from http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf.
Nosek, B. A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. Social Psychology, 45, 137-141.
Orne, M. T. (1962). On The Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications. *American Psychologist*, *17*(11), 776–783.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgement and Decision Making*, *5*(5), 411–419.
Parker, A.M, & Stone, E.R. (2014). Identifying the effects of unjustified confidence versus overconfidence. *Journal of Behavioral Decision Making, 27*, 134.145.
Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., & Jerram, C. (2015). The design of phishing studies: Challenges for researchers. *Computers & Security*. http://doi.org/10.1016/j.cose.2015.02.008
Pattinson, M., Jerram, C., Parsons, K., McCormac, A., & Butavicius, M. (2012). Why do some people manage phishing e-mails better than others? *Information Management & Computer Security*, *20*(1), 18–28.

Perrow, C. (2011). Fukushima and the inevitability of accidents. *Bulletin of the Atomic Scientists*, *67*(6), 44–52.
Ponemon Institute. (2015). *The Cost of Phishing & Value of Employee Training*. Retrieved from https://info.wombatsecurity.com/cost-of-phishing.
Proctor, R. W., & Chen, J. (2015). The Role of Human Factors/Ergonomics in the Science of Security: Decision Making and Action Selection in Cyberspace. Proceedings of *Human Factors: the Journal of the Human Factors and Ergonomics Society*, *57*(5), 721–727. http://doi.org/10.1177/0018720815585906
Purkait, S. (2012). Phishing counter measures and their effectiveness – literature review. *Information Management & Computer Security*, *20*(5), 382–420. http://doi.org/10.1108/09685221211286548
Radicati. (2014). *Email Statistics Report, 2015-2019*. Retrieved from http://www.radicati.com/wp/wp-content/uploads/2015/02/Email-Statistics-Report-2015-2019-Executive-Summary.pdf.
See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-Analysis of the Sensitivity Decrement in Vigilance. *Psychological Bulletin*, *117*(2), 230–249.
Shaw, T. H., Matthews, G., Warm, J. S., Finomore, V. S., Silverman, L., & Costa, P. T. (2010). Journal of Research in Personality. *Journal of Research in Personality*, *44*(3), 297–308. http://doi.org/10.1016/j.jrp.2010.02.007
Sheng, S., Holbrook, M. B., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010). Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In Proceedings of CHI.
Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007). Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish (pp. 1–12). Presented at the Symposium on Usable Privacy and Security (SOUPS), Pittsburgh, PA.
Sloman, S. A. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, *119*(1), 3–22.
Smillie, L. D., Quek, B.-K., & Dalgleish, L. I. (2013). The Impact of Asymmetric Partial Feedback on Response-Bias. *Journal of Behavioral Decision Making*, *27*(2), 157–169.
Smith, A., & Miles, C. (1986). The effects of lunch on cognitive vigilance tasks. *Ergonomics*, *29*(10), 1251–1261. http://doi.org/10.1080/00140138608967238
Sorkin, R. D., & Woods, D. D. (1985). Systems with Human Monitors: A Signal Detection Analysis. *Human–Computer Interaction*, *1*, 49–75.
Stanton, N. A., Booth, N. L., & Stammers, R. B. (1992). Alarms in human supervisory control: a human factors perspective. *International Journal of Computer Integrated Manufacturing*, 5(2), 81–93.
Sun, L., Srivastava, R. P., & Mock, T. J. (2006). An Information Systems Security Risk Assessment Model Under the Dempster-Shafer Theory of Belief Functions. *Journal of Management Information Systems*, *22*(4), 109–142.
Swets, J. A., Dawes, R., & Monahan, J. (2000). Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest*, *1*(1), 1–26.
Symantec Corporation. (2016). *Internet Security Threat Report*. Symantec. Retrieved from https://www.symantec.com/security-center/threat-report

Verizon Communications Inc. (*2016*). *2016 Data Breach Investigations Report.* http://www.verizonenterprise.com/verizon-insights-lab/dbir/2016/
Vishwanath, A. (2015). Examining the Distinct Antecedents of E-Mail Habits and its Influence on the Outcomes of a Phishing Attack. *Journal of Computer-Mediated Communication*, *20*(5), 570–584.
Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, *51*(3), 576–586. doi:10.1016/j.dss.2011.03.002
Wallis, T. S. A., & Horswill, M. S. (2007). Using fuzzy signal detection theory to determine why experienced and trained drivers respond faster than novices in a hazard perception test. *Accident Analysis & Prevention*, *39*(6), 1177–1185. http://doi.org/10.1016/j.aap.2007.03.003
Wang, J., Chen, R., Herath, T., & Rao, H. R. (2009). An Exploration of the Design Features of Phishing Attacks. In Rao and Upadhyaya (Eds.), *Handbooks in Information Systems* (Vol. 4, pp. 259–286). Emerald Group Publishing Limited.
Wang, J. Herath, T., Chen, R., Vishwanath, A., & Rao, H. R. (2012). Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email. *IEEE Transactions on Professional Communication*, *55*(4), 345–362. http://doi.org/10.1109/TPC.2012.2208392
Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance Requires Hard Mental Work and Is Stressful. *Human Factors*, *50*(3), 433–441. doi:10.1518/001872008X312152
Wash, R. (2010). Folk Models of Home Computer Security. Presented at the Symposium on Usable Privacy and Security (SOUPS), Redmond, WA.
Wash, R., & Rader, E. (2015). Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users (pp. 309–325). Presented at the Symposium on Usable Privacy and Security (SOUPS), Ottawa, Canada.
Welch, H.G., Schwartz, L.M., & Woloshin, S. (2011). *Overdiagnosed: Making people sick in the pursuit of health*. Boston: Beacon.
Welk, A. K., Hong, K. W., Zielinska, O. A., Tembe, R., Murphy-Hill, E., & Mayhorn, C. B. (2015). Will the "Phisher-Men" Reel You In? *International Journal of Cyber Behavior, Psychology and Learning*, *5*(4), 1–17. http://doi.org/10.4018/IJCBPL.2015100101
Werlinger, R., Hawkey, K., & Beznosov, K. (2009). An integrated view of human, organizational, and technological challenges of IT security management. *Information Management & Computer Security*, *17*(1), 4–19.
Wickelgren, W. (1977). Speed-accuracy trade-off and information processing dynamics. *Acta Psychologica,* 41(1), 67-85.
Wickens, C.D., Rice, S., Keller, D., Hutchins, S., et al. (2009). False alarms in air traffic control conflict alerting: Is there a "cry wolf" effect? *Human Factors, 51*, 446-462.
Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of Vision*, *13*(3), 33–33. doi:10.1167/13.3.33
Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search

tasks. *Journal of Experimental Psychology: General*, *136*(4), 623–638. http://doi.org/10.1037/0096-3445.136.4.623

Wright R., Chakraborty S., Basoglu A., Marett K. (2010). Where did they go right? Understanding the deception in phishing communications. *Group Decision and Negotiation*, 19, 391–416.

Wright, R. T., & Marett, K. (2010). The Influence of Experiential and Dispositional Factors in Phishing: An Empirical Investigation of the Deceived. *Journal of Management Information Systems*, *27*(1), 273–303. doi:10.2753/MIS0742-1222270111

Wright, R., Chakraborty, S., Basoglu, A., & Marett, K. (2009). Where Did They Go Right? Understanding the Deception in Phishing Communications. *Group Decision and Negotiation*, *19*(4), 391–416. http://doi.org/10.1007/s10726-009-9167-9

Wueest, C. (2014). Targeted Attacks Against the Energy Sector. Symantec. http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/targeted_attacks_against_the_energy_sector.pdf.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. http://doi.org/10.1037/0033-2909.133.5.800

A. Chapter 3 Appendix

The data, R code, and materials for this study are available at https://osf.io/7bx3n/.

A.1. Supporting SDT Analysis

A.1.1. ROC Curves

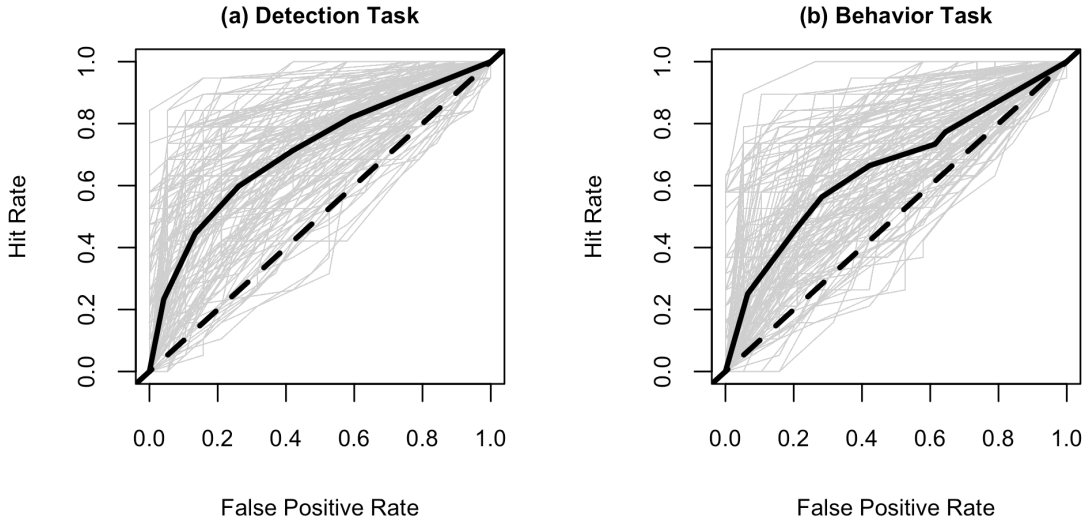

Figure A-1. ROC curves for the (a) detection task and (b) behavior task. The solid black line shows the average ROC curve across all judgments. The grey lines show each individual curve. Performance on both tasks was approximated by an equal-variance Gaussian model.

Responses on the behavior task were scored on an ordinal scale to calculate area under the curve (AUC). The actions, ordered from the most used for legitimate emails to most often used for phishing emails, were: (1) click link or open attachment, (2) reply, (3) ignore or archive, (4) check link, (5) check sender, (6) delete, and (7) report as spam. This order best reflects how participants used the multiple-choice options. Any "other" responses were scored as one of these options based on the explanations that participants provided in the free-text box to calculate d' and c. This ensured that responses from participants who misunderstood the options were correctly interpreted. Any responses that did not fit into the pre-existing categories remained coded as "other". Two coders (one author and one unaffiliated individual) independently coded the items. The inter-rater reliability was 82%. Where the coders disagreed, a third unaffiliated coder independently coded the items and the author-coder reconciled the coding. The table below summarizes the results of the recoding for Experiment 1.

Individuals' performance on the detection and behavior tasks was strongly correlated for AUC, $r(150) = 0.83$, $p < .001$.

Table A-1. Recoding of "other" responses for behavior task.

Action	Times Recoded
Click link or attachment	22
Reply	4
Ignore	5
Other	30
Delete	3
Check sender	97
Check link	217
Report as spam	29
Total	407

A.1.2. Detection vs. Behavior Beta

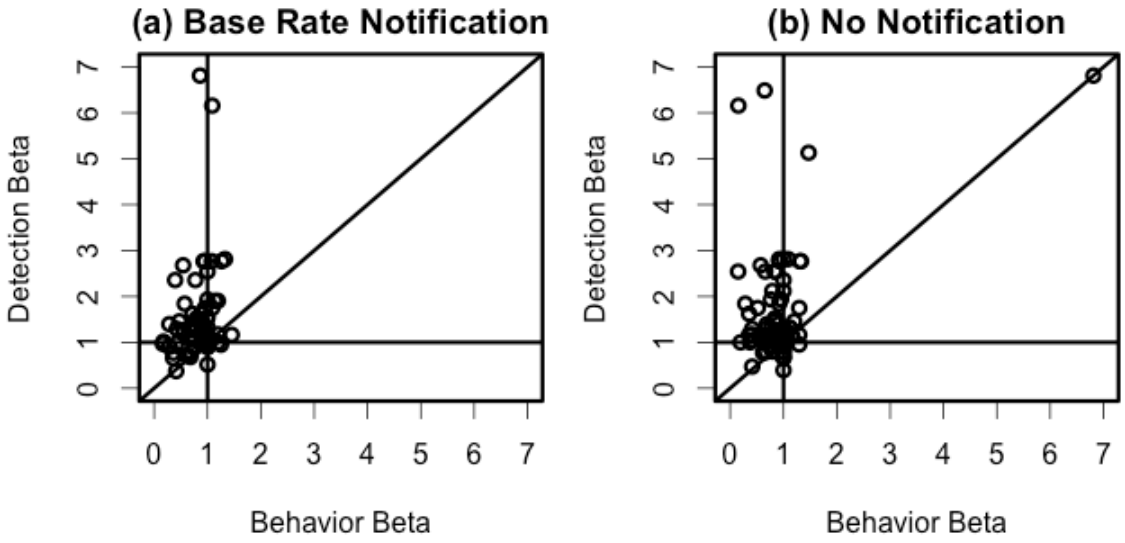

Figure A-2. Behavior vs. detection beta for participants who (a) received the 50% base rate notification and (b) were left to infer the base rate. There is little difference between the base rate notification conditions. Participants tended to have a Behavior Beta < 1, but there was more various for the Detection Beta. Individuals' performance on the detection and behavior tasks was correlated, r(150) = 0.36, p < .001.

A.1.3. Justification of Discrete Choice Model

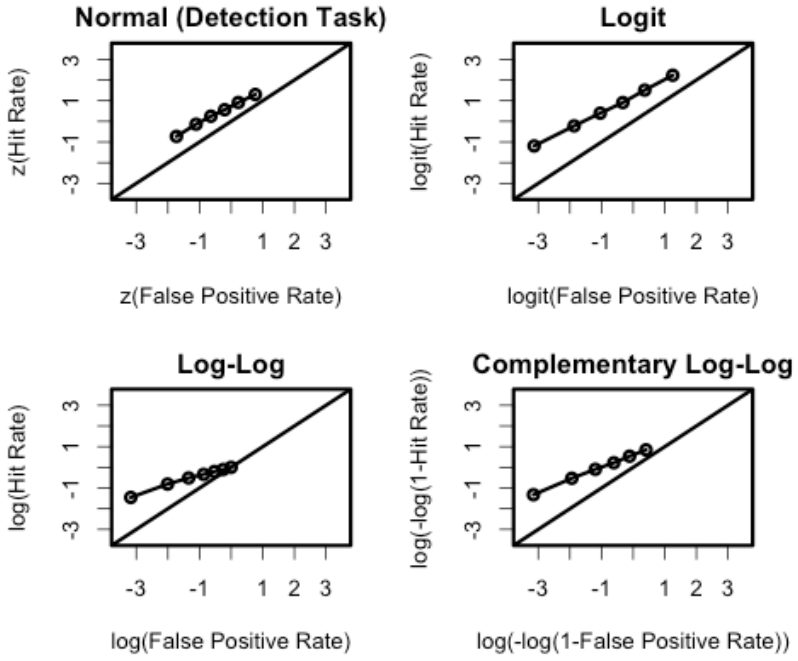

Figure A-3. A comparison of discrete choice models for the detection task. The Normal model is most parallel to the diagonal – suggesting it best fits the data.

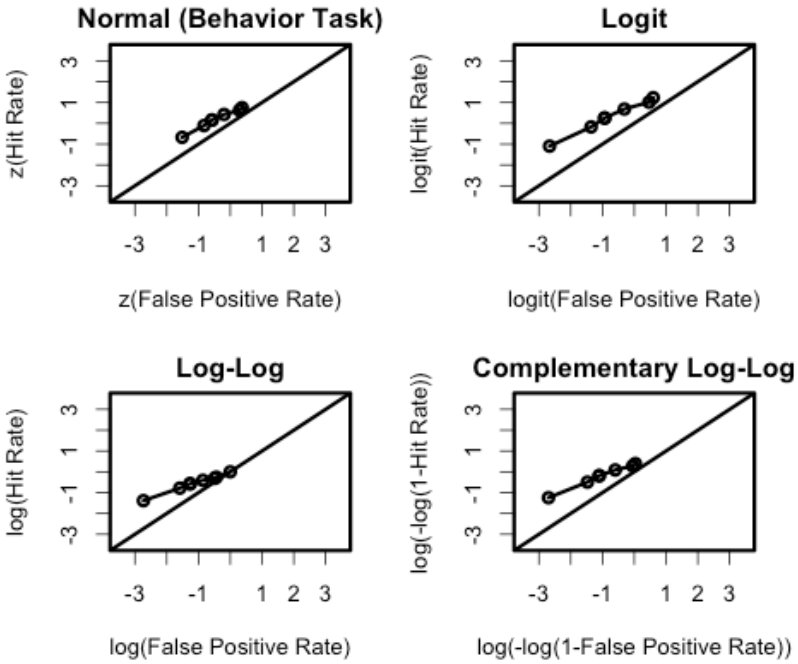

Figure A-4. A comparison of discrete choice models for the behavior task. The Normal model is most parallel to the diagonal – suggesting it best fits the data.

A.2. Supporting Experiment 1 Analysis

A.2.1. Pearson Correlations

Simple Pearson correlations are reported below for the regression analysis.

Table A-2. Pearson correlations for Experiment 1 (N=152).

	1	2	3	4	5
1. Detection d'	1				
2. Detection c	0.01	1			
3. Behavior d'	0.61***	-0.09	1		
4. Behavior c	-0.20*	0.66***	-0.08	1	
5. Attention (1=pass)	0.26**	0.18*	0.12	0.06	1
6. log(Phish info time) (min)	0.18*	0.01	-0.01	-0.08	0.10
7. Median time/email (min)	0.15	-0.02	0.01	-0.32***	-0.01
8. Mean confidence	0.32***	0.28***	0.22**	0.26***	0.09
9. Mean perceived consequences	0.05	-0.35***	0.07	-0.47***	-0.06
10. log(Age)	-0.07	-0.18*	-0.16*	-0.32***	-0.08
11. Gender (1=male)	0.13	-0.05	0.09	0	-0.03
12. College (1=college degree)	0.12	-0.03	0.07	-0.12	-0.01

*p<.05 **p<.01 ***p<.001

	6	7	8	9	10	11
1. Detection d'						
2. Detection c						
3. Behavior d'						
4. Behavior c						
5. Attention (1=pass)						
6. log(Phish info time) (min)	1					
7. Median time/email (min)	0.16	1				
8. Mean confidence	0.06	-0.06	1			
9. Mean perceived consequences	0.11	0.10	-0.04	1		
10. log(Age)	0.08	0.35***	-0.16*	0.21**	1	
11. Gender (1=male)	0	-0.01	0.23**	-0.09	-0.21**	1
12. College (1=college degree)	0.09	-0.02	-0.04	0.09	0.12	-0.03

*p<.05 **p<.01 ***p<.001

Data plots are included below to assess outliers and transformations used in the regression analysis.

A.2.2. Transformations

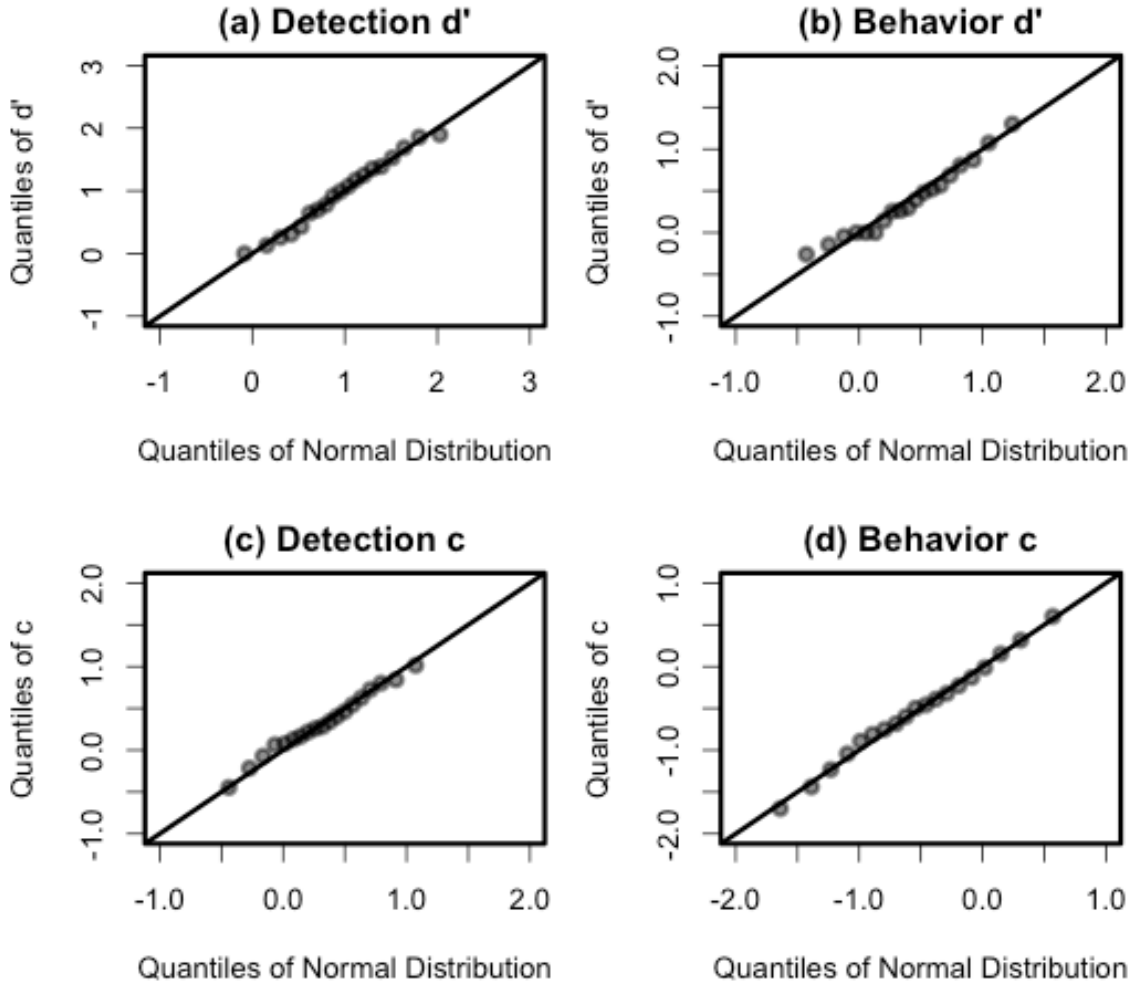

Figure A-5. Q-Q plot of dependent variables. No transformation is needed to assume normality.

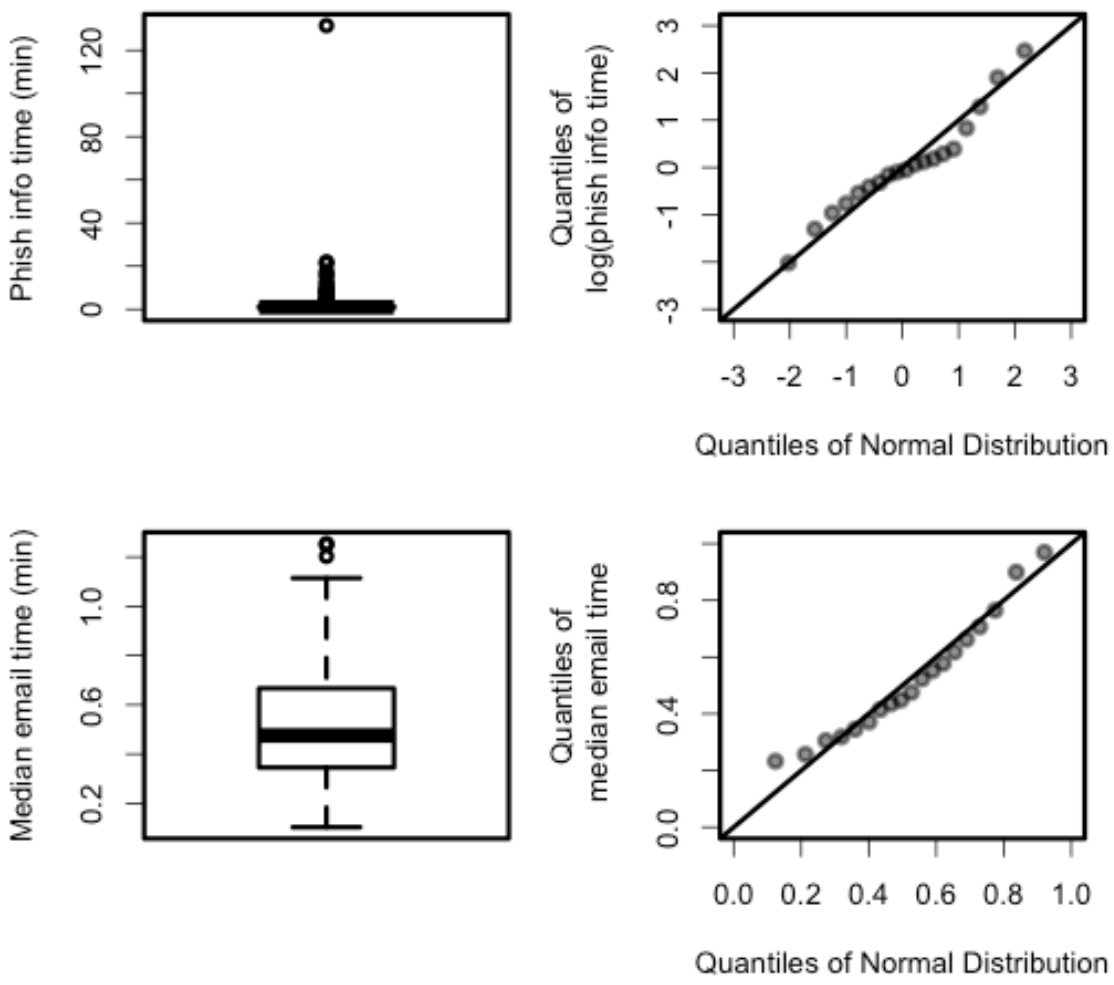

Figure A-6. Boxplots and Q-Q plots for the phish info time and median email time. A log transformation was used for the phish info time due to the high skew and existence of outliers.

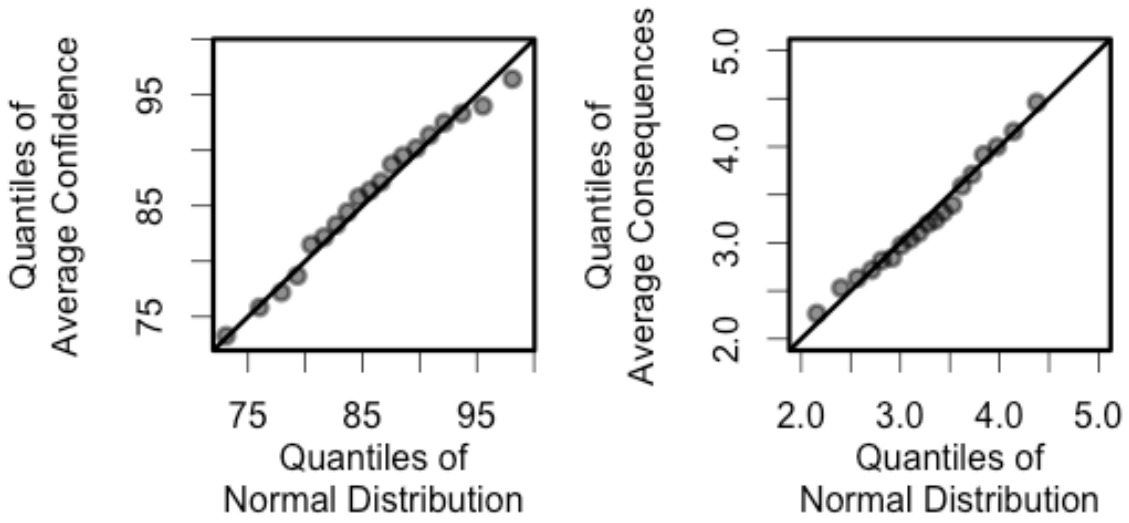

Figure A-7. Q-Q plots for confidence and perceived consequences. No transformations were needed.

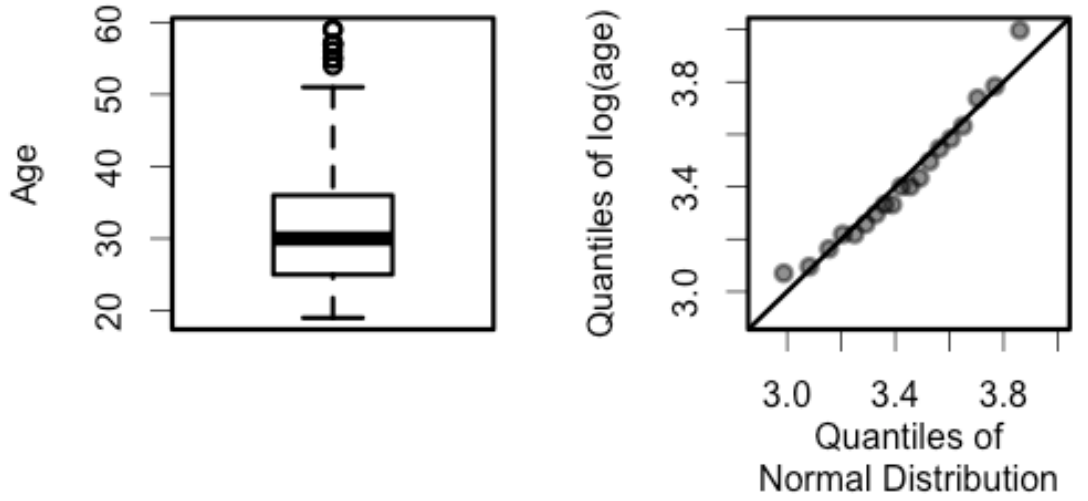

Figure A-8. Boxplot and Q-Q plot for age. A log transformation was used.

A.2.3. Learning Analysis

Table A-3. Detection and behavior d' and c for the first vs. second half of Experiment 1. There were no significant differences, which suggests that no learning occurred.

	Detection Task		Behavior Task	
	First Half M (SD)	Second Half M (SD)	First Half M (SD)	Second Half M (SD)
d'	0.93 (0.77)	0.94 (0.80)	0.40 (0.60)	0.36 (0.69)
c	0.31 (0.51)	0.29 (0.49)	-0.49 (0.66)	-0.51 (0.65)

A.2.4. Alternate Attention Check

We found similar results when excluding the 10 participants who failed 2 of 3 additional attention checks: spending less than 10 seconds on more than one email, illogical responses (e.g., clicking the link on an email identified as phishing) and d' < 0. We would expect attention to occasionally wander so we only excluded participants who failed more than 1 type of test. The estimate of 10 seconds was based on the time it took to scroll down and click responses in each question.

In total, 44 participants spent less than 10 seconds on at least 2 emails, 9 made illogical choices for at least 1 email, and 28 had a negative detection or behavior d'. Of the 10 excluded participants, all failed the timing test (for 1 to 33 messages), 5 made illogical choices, and 7 had negative d'. There were no significant differences in the regression analysis, reported below.

Table A4. Regression analysis excluding participants that failed alternate attention checks.

	d'		c	
	Detection B (SE)	Behavior B (SE)	Detection B (SE)	Behavior B (SE)
Intercept	-1.21 (1.05)	0.09 (0.87)	0.57 (0.72)	0.25 (0.93)
Knowledge of base rate	-0.03 (0.10)	0.07 (0.09)	0.04 (0.07)	0.18 (0.09)
Task order (detection = 1)	0.03 (0.10)	-0.03 (0.09)	-0.01 (0.07)	0.13 (0.09)
Attention (pass = 1)	0.46 (0.22)*	0.11 (0.18)	-0.03 (0.15)	-0.21 (0.19)
log(Phish info time)	0.07 (0.04)	-0.01 (0.04)	-0.01 (0.03)	0 (0.04)
Median time/email	0.39 (0.25)	0.08 (0.21)	0.05 (0.17)	-0.69 (0.22)**
Average confidence	2.37 (0.73)**	1.20 (0.61)	1.34 (0.50)**	2.26 (0.65)***
Average perceived consequences	0.06 (0.08)	0.08 (0.07)	-0.25 (0.06)***	-0.43 (0.07)***
log(Age)	-0.24 (0.21)	-0.35 (0.17)	-0.16 (0.15)	-0.23 (0.19)
Gender (male = 1)	0.10 (0.11)	0.05 (0.09)	-0.09 (0.08)	-0.12 (0.10)
College (college degree = 1)	0.19 (0.11)	0.09 (0.09)	0 (0.07)	-0.15 (0.09)
N	134	134	134	134
Adjusted R^2	0.12	0.03	0.16	0.38
F	2.84**	1.40	3.49***	9.18***

A.3. Supporting Experiment 2 Analysis

Table A-4. Regression models for d' (Experiment 2).

	Detection Task		
	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables B (SE)
Intercept	0.97 (0.16)***	-2.02 (1.38)	-3.46 (2.14)
Knowledge of base rate	0.02 (0.23)	0 (0.24)	0.01 (0.26)
Attention (pass = 1)		0.39 (0.45)	0.50 (0.48)
log(Training time)		0.03 (0.10)	0.02 (0.11)
Median time/email		1.38 (0.77)	0.97 (0.86)
Average confidence		3.21 (1.55)*	3.34 (1.58)*
Average perceived payoffs		-0.17 (0.18)	-0.24 (0.19)
log(Age)			0.49 (0.50)
Gender (male = 1)			-0.14 (0.25)
College (college degree = 1)			-0.01 (0.26)
N	52	47	47
Adjusted R^2	-0.02	0.11	0.07
F	0.01	1.94	1.41

	Behavior Task		
	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables B (SE)
Intercept	0.48 (0.11)***	-1.45 (1.13)	-1.70 (1.72)
Knowledge of base rate	-0.15 (0.16)	-0.18 (0.17)	-0.22 (0.17)
Attention (pass = 1)		-0.25 (0.29)	-0.33 (0.31)
log(Training time)		-0.13 (0.09)	-0.14 (0.09)
Median time/email		0.32 (0.26)	0.45 (0.28)
Average confidence		1.94 (1.16)	2.37 (1.22)
Average perceived payoffs		0.08 (0.13)	0.09 (0.13)
log(Age)			-0.04 (0.38)
Gender (male = 1)			0.20 (0.18)
College (college degree = 1)			-0.23 (0.19)
N	48	47	47
Adjusted R^2	0	0.01	-0.01
F	0.97	1.07	0.93

Table A-5. Regression models for c (Experiment 2).

	Detection Task		
	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables B (SE)
Intercept	0.34 (0.08)***	0.01 (0.76)	-0.48 (1.17)
Knowledge of base rate	-0.08 (0.12)	-0.04 (0.13)	0.01 (0.14)
Attention (pass = 1)		-0.12 (0.25)	-0.12 (0.26)
log(Training time)		-0.03 (0.06)	-0.02 (0.06)
Median time/email		-0.44 (0.42)	-0.57 (0.47)
Average confidence		1.35 (0.85)	1.28 (0.86)
Average perceived payoffs		-0.17 (0.10)	-0.18 (0.11)
log(Age)			0.11 (0.27)
Gender (male = 1)			0.06 (0.13)
College (college degree = 1)			0.18 (0.14)
N	52	47	47
Adjusted R^2	-0.01	0.09	0.07
F	0.42	1.74	1.36

*p<.05 **p<.01 ***p<.001

	Behavior Task		
	Model 1: Task Manipulations B (SE)	Model 2: Stimuli Variables B (SE)	Model 3: Individual Variables B (SE)
Intercept	-0.80 (0.14)***	-0.40 (1.29)	1.14 (1.94)
Knowledge of base rate	0.11 (0.21)	0.08 (0.19)	-0.01 (0.19)
Attention (pass = 1)		-0.66 (0.33)	-0.62 (0.35)
log(Training time)		0.01 (0.10)	0 (0.10)
Median time/email		-0.69 (0.30)*	-0.61 (0.32)
Average confidence		1.64 (1.34)	1.87 (1.38)
Average perceived payoffs		-0.24 (0.14)	-0.24 (0.14)
log(Age)			-0.47 (0.43)
Gender (male = 1)			-0.10 (0.20)
College (college degree = 1)			-0.31 (0.22)
N	48	47	47
Adjusted R^2	-0.02	0.27	0.28
F	0.28	3.77**	2.96**

*p<.05 **p<.01 ***p<.001

A.4. Stimuli

Table A-6. Phishing cues used in stimuli.

Phishing Cue	Example
impersonal greeting	Dear Webmail user
suspicious URL	sonna.com/helpdesk/message-429
unusual communication	I've shared a document with you. It's not an attachment -- it's stored on-line at Google Drive.
request for urgent action	Password will expire in 4 days
grammatical errors or misspellings	View the new mail interface, its easy to access, simply log on with your correct email and password to see new site.

Table A-7. Summary of emails.

Ques.	Phish	Subject	Sender
1	0	eBay Reset Your Password	eBay <ebay@ebay.com>
2	1	IT-Helpdesk Service	IT Help Desk <helpdesk@soma.com>
3	1	Password will expire in 4 days	IT Help Desk <helpdesk@soma.com>
4	0	Soma Account Deactivation	IT Help Desk <helpdesk@soma.com>
5	1	Access New Interface	IT Help Desk <helpdesk@soma.com>
6	1	Alumni Panel?	Karen Laurel <karen.laurel@gmail.com>
7	1	Attention	Tom Daniels <tdaniels@soma.com>
8	1	Customer Alert	Capital One <capitalone@gmail.com>
9	0	Activate a new feature in your account	Paypal <paypal@e.paypal.com>
10	0	Annual Report	Mark Hous (via Google Drive) <mhous@gmail.com>
11	0	Data Tracking Article	Ben Farm <bfarm@soma.com>
12	0	Google Apps @ Soma Storage Increase	IT <it@soma.com>
13	1	Cyber Security Awareness Month: Take Security 101	Mary Ann Bane <mabane@soma.com>
14	1	Double Frequent Flyer Miles!	Customer Appreciation <cust@boa.com>
15	1	Invitation to connect on Linkedin	Linkedin <member@linkedin.com>
16	0	Important – eBay Password Reset Required	eBay <eBay@reply1.ebay.com>
17	0	IMPORTANT MESSAGE FROM HEALTH SERVICES	Health Services <hs-noreply@soma.com>
18	0	Important Phishing Notice – Please Read	Mary Ann Bane <mabane@soma.com>
19	0	Kelly, people are looking at your Linkedin profile	Linkedin <messages-noreply@linkedin.com>
20	0	New voicemail from (724) 970-8435 at 12:27 PM	Google Voice <voice-noreply@google.com>
21	1	(no subject)	Carlos Saborio Villalta <carlossaboriovillalta@gmail.com>
22	1	SomaTRAK Update	HR <hr@soma.com>
23	0	Scanned Document From PRINT4.SOMA.COM	PRINT4-SOMA <print4@soma.com>

24	0	TurboTax Notice: Your Privacy Statement	TurboTax Team <TurboTax@turbotax.intuit.com>
25	0	Update your Business Info with SCHS	Peggy Bittner <bittner@schs.org>
26	0	UPS Ship Notification, Tracking Number 1Z4531280357253423	UPS Quantum View <auto-notify@ups.com>
27	0	USAID Job Openings	Jenna Martin <jennam@gmail.com>
28	1	UNICSID VACANCY NEWSLETTER	George Lancy <George.lancy@un-icsid.org>
29	1	UPS Shipment Authorization	UPS <auto-notify@ups.com>
30	1	*** Urgent Notification ***-499348210	Security <security@bankofireland.com>
31	1	Webmail Alert Notice	IT <it@soma.com>
32	1	Weekly Meeting Agenda – URGENT	Mark Hous (via Google Drive) <mhous@gmail.com>
33	0	WorldPay CARD transaction Confirmation	Shopper <shopper@worldpay.com>
34	0	Your credit card is about to expire	Netflix <info@mailer.netflix.com>
35	0	Your receipt No.130086326136	iTunes Store <do-not-reply@itunes.com>
36	1	Your Apple ID was disabled	Apple <accounts@apple.com>
37	1	Your Email Account	IT Help Desk <helpdesk@soma.com>
38	1	Your Salary Raise Confirmation	Matt Henn <mhenn@soma.com>

B. Chapter 4 Appendix

The data and R code for this study are available at https://osf.io/hwjmn/.

B.1. Preregistration Document[2]

1. INTRODUCTION

We have demonstrated a method for measuring users' vulnerability to phishing attacks in a scenario-based online experiment (Canfield, Fischhoff & Davis, 2016; following Kumaraguru et al., 2010). Specifically, we used signal detection theory (SDT) to disentangle users' ability to distinguish between phishing and legitimate emails (discrimination ability or d') and tendency to classify emails as phishing or legitimate (decision threshold or c) for both detection and behavior decisions. Users with a high d' are better able to distinguish between phishing and legitimate emails. Users with a positive c are biased toward believing emails are legitimate, while the inverse is true for negative c. This work suggests that although users tend to choose cautious behaviors, they aren't able to sufficiently compensate for poor detection ability. However, this work is limited to a laboratory environment and lacks the complexity of real world data. Thus it is unclear to what degree these parameters explain actual, rather than theoretical, phishing risk.

The Security Behavior Observatory (SBO) is an existing effort to gather ecologically valid data on users' security habits over time (Forget et al., 2014). Participants agree to have software installed on their personal computers, which collects data on browsing, installed applications, processes, network connections, and events. We propose comparing performance on our phishing detection experiment to behavioral measures from the SBO. This will serve to validate our measurement of phishing risk.

1.1. AIMS

We have two primary aims for this study:

1. Replicate Chapter 3 with a non-mTurk sample.
2. Assess construct validity of Chapter 3 SDT measures.

---

[2] This document details my data analysis approach before I combined the SBO and experimental data sets. This serves to clearly distinguish hypothesis generation from hypothesis testing (Miguel et al., 2014; Nosek & Lakens, 2014). Sarah Pearman performed the mapping between the SBO and experimental data sets. I had no access to the information (participant emails) needed to combine the datasets. Once the experimental data set was added to the SBO database, the mapping was added. This occurred post preregistration (3/29/2016). The preregistration is at the Open Science Framework (osf.io/uagmc) and currently under embargo.

We expect that users who have more negative behavioral outcomes (e.g. more malware, more visits to phishing domains) will have a lower d' and higher c for both detection and behavior tasks. We will test this by looking at the simple association between these behavioral outcomes and the SDT measures, as well as how these SDT measures improve the fit of a multiple regression model predicting those outcomes.

2. METHOD

2.1. SAMPLE

We recruited participants from the Security Behavior Observatory (SBO) study (Forget et al., 2014) to perform a phishing detection experiment. The SBO participants are primarily college students and retired people in the Pittsburgh area. Participants opt-in to the SBO study and are aware that their computer use is being monitored. Therefore, we should address concerns about opt-in bias and Hawthorne Effect. Each participant was paid $20 to complete the phishing detection experiment.

2.2. DESIGN

To replicate Chapter 3, SBO participants performed the phishing detection experiment. Participants reviewed emails of a fictitious person to judge whether or not each email was phishing (detection task) and what action they would perform on the email (behavior task). No changes were made to the experimental design, but the individual difference questions asked at the end of the experiment were modified. For each participant, we used SDT to estimate d' and c.

To assess construct validity, we used the SDT estimates to predict negative real world outcomes. For each SBO participant, we assessed 2 negative outcomes, existence of malicious software and visits to malicious websites. For each outcome, we first measure the simple relationship between the outcome and the SDT measure using logistic regression. Then we use a likelihood ratio test to compare the goodness of fit for models with and without the SDT measures.

2.3. MEASURES

2.3.1. PHISHING DETECTION MEASURES

We have five phishing detection performance measures from the experiment:

1. Detection d': measure of ability to tell the difference between phishing and legitimate emails.
2. Detection c: measure of bias toward identifying an email as phishing (negative c) or legitimate (positive c).
3. Behavior d': measure of ability to distinguish between when to click on links and when not to.
4. Behavior c: measure of bias toward clicking on links (positive c) or not (negative c).
5. λ or P(click|"legit"): percent of times that users click on links that they perceive as legitimate. This is a potential alternative to using behavior d' and behavior c in risk analysis.

We changed the individual difference measures from Chapter 3 to account for additional sources of variance. Table 1 describes potential covariates from these data. If significantly related, these variables will be added to a second stage of validation models.

Table B-1. Self-reported covariates from phishing detection experiment.

Covariate	Definition	Hypothesis
Age of computer	Reported in years	+ Users who have older computers have more malware.
Total email load	Average emails per day	+ Users who get more emails have visited more phishing websites.
Self-reported behavior	Frequency of email, social media, instant messenger, software downloads	+ Users who report being more active have visited more phishing websites and have more malware.

2.3.2. SBO MEASURES

The SBO measures are based on behavior recorded by the SBO software on participants' personal computers. These measures only record activity on a particular computer, which may not reflect all of an individual's activity (e.g. if they have a separate work computer).

For another study using SBO data, participants were asked to complete the Security Behavior Intention Scale (Egelman & Peer, 2015). The SeBIS has four subscales including device securement, password generation, proactive awareness, and updating. We included this self-reported scale in our replication and validation analysis.

2.3.2.1. BROWSING VARIABLES

To assess browsing, we used data from 2 separate sensors, browsers (Internet Explorer, Chrome, and Firefox) and network packets. Browser sensors capture web browsing. The network packet sensor captures all http traffic, which includes ads and images separately from websites.

The outcome variable for browsing is Phish Visits, which is a count of how many known phishing domains are visited on each user's computer. Known phishing domains were identified using the Google Safe Browsing dataset. This is a count rather than normalized across time (e.g. phish visits/month) to facilitate interpretation. For example, for phish visits/month, it is difficult to account for periods of high vs. low usage (e.g. the start of the school year or if someone goes on vacation).

Potential browsing covariates are summarized in Table 2. In addition, we expect both Click Email Links to be positively correlated with λ, meaning that users who

click on links from email will click on more links that they perceive to be legitimate in the experiment.

Table B-2. Browsing Covariates.

Covariate	Definition	Hypothesis	
Days in Study	max(Sensor Time) – min(Sensor Time)	+	Users who have been in the study longer have visited more phishing websites.
Total Browsing	Count of URLs visited	+	Users who have visited more websites have visited more phishing websites.
Unique Browsing	Count of unique URLs visited (excluding URLs that were visited multiple times)	+	Users who have visited more unique websites have visited more phishing websites.
Unique Domains	Count of unique domains visited	+	Users who have visited more unique domains have visited more phishing websites.
Social Media Use	Count of social media domains visited: Facebook, Twitter, LinkedIn, Google+, Tumblr, Pinterest, Instagram, and Reddit	+	Users who visit more social media domains have visited more phishing websites.
Click Email Links (track)	Count of links clicked from email. This count URLs that have built-in tracking to determine that they came from email (i.e. links that include “mail” or “email” after =, &, or ?, excluding email domains).	+	Users who click on links from email that use tracking are more likely to have visited phishing websites.
Click Email Links (source)	Count of links clicked from webmail. This counts URLs where the source URL is an email domain and the destination is not. This does not capture links clicked from an email client (e.g. Outlook).	+	Users who click on links from webmail are more likely to have visited phishing websites.

Note: Where possible, each variable was estimated from both the browser and network packet sensors.

2.3.2.2. SOFTWARE VARIABLES

The outcome variable for software is malicious software. We used two measures to distinguish between malicious and suspicious software. Malware is a count of

known malicious programs on each user's computer. Suspicious Software is a count of suspicious programs on each user's computer. All malware is suspicious but not all suspicious programs are malware. Suspicious programs have a broader definition that includes adware.

Each piece of installed software was coded by hand to indicate if it was malicious or suspicious (up to October 2015 data) using shouldiremoveit.com. At present, only 43% of the observed software has been coded. We are currently developing an automated process for identifying malware and suspicious software. Once complete, we expect to have more observations of malware and suspicious software.

Potential software covariates are summarized in Table 3. Although it is unintuitive, we expect security software to be positively correlated with malware for several reasons. First, more security software does not make a computer more secure. For example, installing more than one third-party anti-virus can reduce the effectiveness because of conflicts between the anti-virus programs. Second, some security software is actually malware. Third, users who have more malware in the first place may choose to install more security software in an attempt to fix their computer. As a result, security software may be reactive, rather than proactive.

Table B-3. Software Covariates.

Covariate	Definition	Hypothesis
Total Software	Count of installed software	+ Users who have more installed software have more malware.
Security Software	Count of security software. Security software includes anti-virus, password managers, security-related browser toolbars, and anti-theft programs.	+ Users who install more security software have more malware.
3rd Party AV	Binary indicator of single 3rd party anti-virus installed (recommended best practice)	- Users with single 3rd party AV have less malware.
Delayed Software Updates	Count of outdated versions of popular software: Adobe Flash, Adobe Reader, Java, Internet Explorer, Chrome, and Firefox	+ Users who delay updating popular software have more malware.
Days Since Windows Update	Days since Windows update was installed. This does not capture why users waited to install updates (e.g. users who delayed updates vs. had not been prompted yet if computer was off).	+ Users who delay updating Windows have more malware.

2.4. ANALYSIS

In this section, we distinguish between the analysis used for (1) replicating Chapter 3, (2) building a model from the SBO data, and (3) assessing construct validity.

2.4.1. REPLICATION

We repeated the SDT calculations and regression analysis from Chapter 3. For each individual, we estimated the SDT parameters by assuming the signal and noise distributions were Gaussian with equal variance and using a log-linear correction (Hautus, 1995; Lynn & Barrett, 2014):

$$d' = z(H) - z(FA)$$
$$c = -0.5(z(H) + z(FA))$$

where

$$H = (\text{hits} + 0.5)/(\text{signals} + 1)$$
$$FA = (\text{false alarms} + 0.5)/(\text{noise} + 1)$$

We assessed the effect of personal greeting using a multilevel model and tested the other hypotheses using multivariate linear regression. In this study, we added total email load and SeBIS scores to the regression.

2.4.2. SBO MODEL CONSTRUCTION

In order to assess construct validity, we constructed models for the browsing and software outcomes. We constructed the models using all of the available SBO data and then refined them with the subset that performed the phishing detection experiment. We used the strategy outlined in Hosmer, Lemeshow & Sturdivant (2013). This involves the following steps:

1. Perform univariate analysis of each IV using a chi-square test for categorical variables and two-sample t-test for continuous variables. All variables with a $p < .25$ are identified as viable covariates. The threshold of .25 was determined by Mickey and Greenland (1989), who show that this threshold ensures that no important variables are eliminated at this initial step.
2. Perform factor analysis to assess combining variables.
3. Identify necessary transformations and link function for linear relationships using GAM and Stukel's test.
4. Fit multivariate model and eliminate variables that are not significant. Test improved model fit using likelihood ratio test. Check that none of the coefficients have dramatically changed.
5. If relevant, check for significant interactions.
6. Assess calibration of final model.

2.4.3. VALIDITY

We evaluated 4 SBO outcomes including (1) phish visits in browser data, (2) phish visits in network packet data, (3) installed malware, and (4) installed suspicious software. For each SBO outcome variable, we used a logistic regression model due to the high number of participants who have a 0 outcome (i.e. have never visited a malicious website and have no malware) (Long, 1997). For each outcome, we measured the simple relationship with the SDT measure. Then we used a likelihood

ratio test to compare the goodness of fit for models with and without the SDT measures. The likelihood ratio test is the most powerful test of the null hypothesis that the SDT measure does not increase the likelihood of the data given the SDT measure.

3. RESULTS

Results are reported where the SBO and experimental data sets were analyzed separately. The combined analysis will be performed post-preregistration. The SBO data changes over time as more data is collected. The final data analysis will be preformed on the most recent database.

3.1. SAMPLE

We recruited 132 participants to participate in the phishing detection experiment. Of those, 121 participants started the survey and 98 finished, giving a 74% response rate. According to two-sample t-tests, the SBO sample is significantly more educated, $t(214) = 3.16$, $p = 0.002$, and older, $t(130) = 4.32$, $p < 0.001$, than the mTurk sample.

Of the 98 participants, 71 failed at least one attention check. Of those, 12 participants failed one of the direction checks and 69 failed one or both of the email stimuli checks. However, no participants were removed for performance on the attention checks because attention was not a significant predictor in the regression analysis. This means that there were no significant differences in performance between participants who did and did not fail the attention checks, which suggests that the checks did not measure attention. Therefore, we are investigating other variables, such as time per stimuli and patterns of responses to assess attention.

Table B-4. Comparison of mTurk and SBO samples.

Variable	mTurk Sample	SBO Sample	All SBO
Gender	58% Female	60% Female	61% Female
Education	45% Bachelors+	64% Bachelors+	58% Bachelors+
Age	32 [19, 59]	40 [19, 81]	46 [19, 87]
N	151	98	213

3.2. REPLICATION

Not reported in preregistration.

3.3. SBO MODEL CONSTRUCTION

3.3.1. BROWSING MODEL CONSTRUCTION

In the browsing data, 7 participants visited a malicious domain in the browser data and 22 participants visited a malicious domain in the network packet data. The outcomes of the two sensors are weakly correlated so it does not make sense to combine them. Therefore, we performed separate regressions for each sensor.

Study Time was not correlated with phishing visits for the browser and network packet data. Therefore it was excluded from the analysis. We were unable to perform univariate analysis due to the low number of observed phishing visits. According to the factor analysis reported in Table 5, all of the remaining browsing covariates load on a single factor called Browsing Intensity. The logistic regression is reported in Table 6. Browsing Intensity was a significant predictor for the network packet sensor data, but not browser data. Figure 1 shows the predicted probability of having visited a malicious URL, given the browsing intensity.

Table B-5. Factor analysis for browsing variables. The factor analysis is reported separately for the browser and network packet sensors. Browsing Intensity is a linear combination of all of the browsing variables for each sensor.

	Browser Sensor	**Network Packet Sensor**
Unique Browsing	0.99	0.91
Total Browsing	0.97	0.84
Click Links (track)	0.89	0.32
Unique Domains	0.87	0.94
Social Media Use	0.51	0.64
Click Links (source)	0.26	N/A
% of Total Variance	0.63	0.59
Cronbach's Alpha	0.89	0.84

Table B-6. Logistic model for browsing variables. Browsing intensity requires a log transformation to meet the assumption of linear parameters. Browsing intensity is not a significant predictor for the browser sensor, likely due to insufficient observations of phish visits. The odds ratio measures the change in the odds of the outcome from a 1-unit change in the predictor. For the network packet data, a 1-unit increase in log(Browsing Intensity) increases the odds of visiting a phishing URL by 1.23 times.

	Browser Sensor		Network Packet Sensor	
	B (SE)	OR	B (SE)	OR
(Intercept)	-5.7 (2.1)**		-8.84 (2.47)***	
log(Browsing Intensity)	0.43 (0.24)	1.27	0.66 (0.21)**	1.23
N	64		96	

Note: B = Beta (regression coefficient), SE = Standard error, OR = Odds ratio
*p < .05, **p<.01, ***p<.001

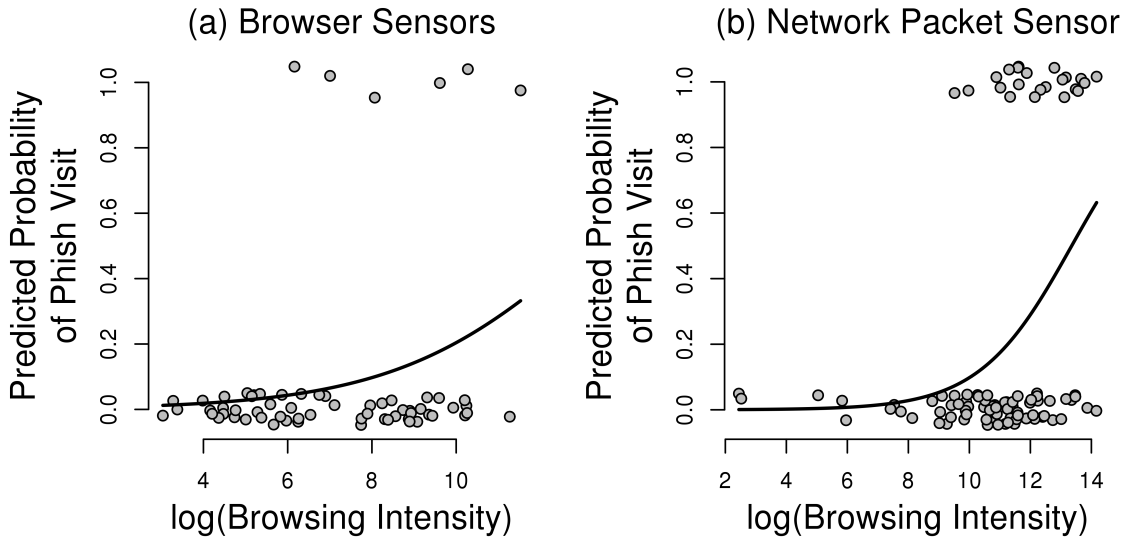

Figure B-1. Predicted probability of visiting a phishing website (line) plotted on top of observations (points) vs. log(Browsing Intensity).

3.3.2. SOFTWARE MODEL CONSTRUCTION

In the software data, 26 participants had malware and 44 had suspicious software installed on their computer. Since malware is a subset of suspicious software, it is more enlightening to analyze them separately.

Univariate and multivariate analysis indicated that Days Since Windows Update and 3rd Party AV did not predict whether or not users had malware or suspicious software on their computer. Therefore, those variables were excluded from the logistic model.

According to the factor analysis reported in Table 7, it is appropriate to combine the three remaining variables, Total Software, Security Software, and Delayed Software Updates, into a single factor called Software Load. Participants with high Software Load tended to have more software installed on their computer, including more security software, and delay software updates on popular software. In the logistic regression analysis reported in Table 8, participants with a higher software load were significantly more likely to have malware and suspicious software on their computer. Figure 2 shows the predicted probability of having malware or suspicious software, given the software load.

Table B-7. Factor analysis for software variables. These variables were combined to form the Software Load variable.

	Factor 1
Total Software	0.69
Delayed Software Updates	0.56
Security Software	0.49
% Variance	0.34
Cronbach's Alpha	0.6

Table B-8. Logistic regression of malware and suspicious software. The odds ratio measures the change in the odds of the outcome from a 1-unit change in the predictor. Here, Software Load is scaled so that 1-unit = 10-units. An odds ratio of 1 suggests that there is little effect from a 10-unit increase in Software Load. However, the effect is multiplicative so a 100-unit increase in Software Load increases the odds of having malware or suspicious software by 2.72.

	Malware		Suspicious Software	
	B (SE)	OR	B (SE)	OR
(Intercept)	-2.27 (0.46)***		-1.56 (0.43)***	
Software Load (by 10s)	0.1 (0.03)***	1.03	0.12 (0.03)***	1.03
N	96		96	

Note: B = Beta (regression coefficient), SE = Standard error, OR = Odds ratio
*p < .05, **p<.01, ***p<.001

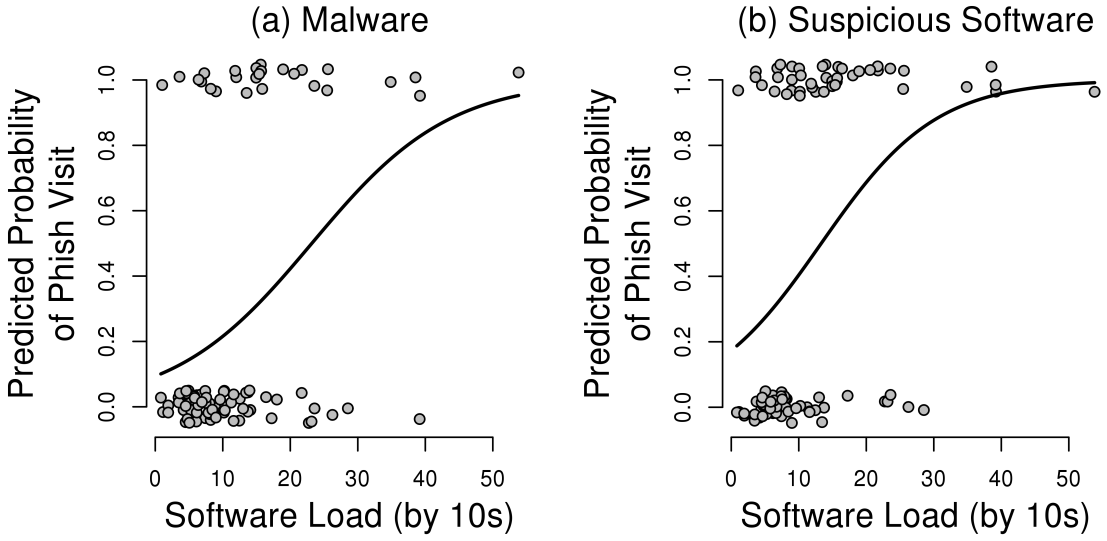

Figure B-2. Predicted probability of having malware or suspicious software (line) plotted on top of observations (points) vs. Software Load.

3.4. VALIDITY

To assess whether the SDT parameters improve the fit of the models, we will perform a likelihood ratio test comparing the models in Table 9 with the nested models shown in Tables 6 and 8. We will compare the results for d' with AUC (area under the curve), an alternate discrimination measure based on the ROC curve.

Table B-9. Example logistic regression for SBO and SDT data. This regression will be repeated for each of the four outcome variables, (1) phish visits in browser data, (2) phish visits in network packet data, (3) installed malware, and (4) installed suspicious software. First, we will measure the simple relationship for each model by excluding Factor 1. Then we will perform the models described below. Factor 1 will be log(browsing intensity) for the network packet data and software load for the software data. There is no Factor 1 for the browser data.

	Model 1: d'_D	Model 2: c_D	Model 3: d'_B	Model 4: c_B	Model 5: λ	Model 6: Full	Model 7: Full (λ)
Intercept	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)
d'_D	B (SE)					B (SE)	B (SE)
c_D		B (SE)				B (SE)	B (SE)
d'_B			B (SE)			B (SE)	
c_B				B (SE)		B (SE)	
λ					B (SE)		B (SE)
(Factor 1)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)
N	X	X	X	X	X	X	X

Note: This analysis will be performed post-preregistration
B = beta, the estimated coefficient, SE = standard error of beta
Alpha = .01

4. DISCUSSION

If none of the SDT parameters predict the SBO outcomes, there are four primary explanations:

1. SDT parameters do not measure phishing susceptibility: participants may pay attention to different cues in the experiment than in real life because of the constrained context and training. We can evaluate whether training time is significant (it was not for the mTurk sample), use total email as a predictor (since people who get more email may be more rushed or more knowledgeable), and code responses on how they identify phishing emails in real life vs. in the experiment.
2. SDT measurement is imprecise: participants may not be good at the experimental task or not try hard because it is unusual, tedious, or boring. We can test for vigilance decrement (decreased performance over time) and use the attention checks.
3. SBO outcomes are not related to phishing susceptibility: the behavioral outcomes may be primarily from other vectors (e.g. downloading software, social media, instant message) or participants may fall for phishing attacks

on other devices (e.g. phone). In the experiment, we ask about the frequency of these activities.

4. SBO measurement is imprecise: the data may be missing (e.g. sensor breaks or is turned off, participant primarily uses different computer) or noisy (multiple people using the same computer). In the experiment, we ask if participants have turned off browser extensions and if other people use their computer.

If the SDT parameters predict phish visits, this suggests that the experiment is truly measuring people's phishing susceptibility. If the SDT parameters predict malware, this may imply that the experiment is measuring people's ability to detect suspiciousness on the Internet in general, rather than being limited to detecting phishing emails.

B.2. Supporting Analysis

Table B-10. Comparison of linear regression analysis of sensitivity (d') for mTurk and community samples.

	mTurk		Community (SBO)	
	Detection	Behavior	Detection	Behavior
	B (SE)	B (SE)	B (SE)	B (SE)
Intercept	-1.32 (0.98)	-0.09 (0.83)	-0.97 (0.92)	0.61 (0.77)
Knowledge of base rate	0.02 (0.10)	0.10 (0.08)	-0.22 (0.14)	0.05 (0.12)
Task order (detection = 1)	0.04 (0.10)	-0.05 (0.09)	0.15 (0.14)	-0.14 (0.12)
Attention (pass = 1)	0.49 (0.18)**	0.12 (0.15)	0.33 (0.19)	0.07 (0.16)
log(Phish info time)	0.05 (0.04)	-0.03 (0.03)	0.02 (0.07)	0.07 (0.06)
Median time/email	0.48 (0.23)*	0.17 (0.19)	0.23 (0.11)*	-0.13 (0.09)
Average confidence	2.23 (0.67)**	1.11 (0.57)	3.46 (0.87)***	0.71 (0.73)
Average perceived consequences	0.08 (0.08)	0.11 (0.06)	0 (0.10)	0 (0.08)
log(Age)	-0.22 (0.21)	-0.33 (0.17)	-0.40 (0.19)*	-0.16 (0.16)
Gender (male = 1)	0.11 (0.10)	0.06 (0.09)	-0.09 (0.15)	0.15 (0.12)
College (college degree = 1)	0.19 (0.10)	0.10 (0.09)	-0.03 (0.16)	-0.18 (0.13)
N	142	142	84	84
Adjusted R^2	0.16	0.05	0.14	0.05
F	3.71***	1.68	2.37*	1.40

Table B-11. Comparison of linear regression analysis of response bias (c) for mTurk and community samples.

	mTurk		Community (SBO)	
	Detection	Behavior	Detection	Behavior
	B (SE)	B (SE)	B (SE)	B (SE)
Intercept	0.06 (0.70)	0.10 (0.87)	1.31 (0.62)*	-0.08 (0.81)
Knowledge of base rate	0.01 (0.07)	0.13 (0.09)	0 (0.10)	0.07 (0.13)
Task order (detection = 1)	-0.01 (0.07)	0.11 (0.09)	0.18 (0.10)	0.12 (0.13)
Attention (pass = 1)	0.08 (0.13)	-0.19 (0.16)	0.07 (0.13)	-0.13 (0.17)
log(Phish info time)	0.01 (0.03)	0.01 (0.04)	0.04 (0.05)	0 (0.06)
Median time/email	0.10 (0.16)	-0.70 (0.20)***	0.13 (0.08)	-0.10 (0.10)
Average confidence	1.81 (0.48)***	2.38 (0.59)***	0.62 (0.59)	0.93 (0.77)
Average perceived consequences	-0.24 (0.05)***	-0.42 (0.07)***	-0.24 (0.07)***	-0.20 (0.09)*
log(Age)	-0.17 (0.15)	-0.22 (0.18)	-0.38 (0.13)**	-0.18 (0.16)
Gender (male = 1)	-0.13 (0.07)	-0.14 (0.09)	0.05 (0.10)	0.11 (0.13)
College (college degree = 1)	0.02 (0.07)	-0.13 (0.09)	0.39 (0.11)***	0.18 (0.14)
N	142	142	84	84
Adjusted R^2	0.18	0.39	0.27	0.07
F	4.16***	9.85***	4.12***	1.63

Table B-12. Descriptive statistics for validity analysis.

	Mean	SD	Med	Min	Max	N
SeBIS	3.3	0.52	3	2	5	83
SeBIS – Device	3.14	1.28	3	1	5	83
SeBIS – Password	3.05	0.79	3	1	5	83
SeBIS – Proactive Awareness	3.53	0.66	4	2	5	83
SeBIS – Update	3.44	0.82	3	1	5	83
Days in SBO Study	230	170	180	15	658	93
Active Days (browser)	67	76	40	1	438	86
Total URLs/Day (browser)	56	90	22	0	531	68
Unique URLs/Day (browser)	23	32	9	0	151	68
Domains/Day (browser)	6	4.4	5	1	32	86
Clicked Email Links/Day via tracking (browser)	0.5	1	0	0	5	68
Clicked Email Links/Day via source (browser)	0.8	2	0	0	13	68
Active Days (network packet)	85	63	70	1	347	92
Total URLs/Day (network packet)	2,600	3600	1500	6	27,225	92
Unique URLs/Day (network packet)	990	1000	670	4	5,215	92
Domains/Day (network packet)	52	37	42	3	216	92
Clicked Email Links/Day via tracking (network packet)	0.4	2	0	0	19	92
Total Software	342	320	240	15	1,573	92
Vulnerable Software	2	1.2	2	0	5	93
AV (binary)	0.34	0.48	0	0	1	93
Last Windows Update (days)	59	34	71	0	180	93

Table B-13. Pearson correlations.

	1	2	3	4	5	6	7	8	9	10
1. Malicious URL (br)	1	0.23	0.11	-0.06	0.04	-0.17	0.07	-0.12	0.19	-0.03
2. Malicious URL (NP)	0.23	1	0.05	0.08	-0.02	-0.04	0.07	0.05	0.06	-0.11
3. Malware	0.11	0.05	1	0.16	-0.15	-0.13	-0.05	-0.06	-0.13	-0.02
4. Malicious Files	-0.06	0.08	0.16	1	-0.1	-0.17	-0.08	-0.19	-0.07	0.14
5. Detection d'	0.04	-0.02	-0.15	-0.1	1	0.23	**0.53**	0.03	**0.31**	0
6. Detection c	-0.17	-0.04	-0.13	-0.17	0.23	1	0.02	**0.6**	-0.03	**-0.43**
7. Behavior d'	0.07	0.07	-0.05	-0.08	**0.53**	0.02	1	0.17	0.07	-0.01
8. Behavior c	-0.12	0.05	-0.06	-0.19	0.03	**0.6**	0.17	1	0.07	**-0.3**
9. Confidence	0.19	0.06	-0.13	-0.07	**0.31**	-0.03	0.07	0.07	1	0.2
10. Perceived Consequences	-0.03	-0.11	-0.02	0.14	0	**-0.43**	-0.01	**-0.3**	0.2	1
11. Male	0.13	**0.31**	0.01	-0.18	-0.02	0.15	0.14	0.19	0.05	-0.2
12. Age	-0.08	-0.13	0.17	-0.01	-0.15	-0.16	-0.21	-0.16	0.11	0.15
13. College	-0.14	-0.12	0.09	-0.17	-0.08	0.21	-0.19	0.02	0.07	-0.02
14. SeBIS	-0.27	0	-0.05	0.07	0.02	-0.2	0.02	-0.26	0.11	**0.34**
15. Device Subscale	-0.23	0.03	-0.06	0.11	-0.11	-0.12	-0.02	0.03	-0.01	0.24
16. Password Subscale	**-0.31**	-0.13	0.07	-0.05	0.02	-0.13	-0.14	-0.25	-0.11	0.2
17. Proactive Awareness Subscale	-0.07	0.02	-0.02	0.01	0.1	-0.07	0.13	**-0.28**	0.16	0.18
18. Update Subscale	0.1	0.09	-0.09	0.06	0.13	-0.18	0.09	-0.25	**0.3**	0.15
19. Time in Study	-0.08	-0.01	**0.32**	0.21	-0.13	-0.15	-0.12	-0.07	0.07	0.02
20. Active Days (br)	0.26	0.16	0.25	0.04	-0.03	-0.29	0.11	-0.18	0.06	0.07
21. Total URLs/Day (br)	0.18	0.11	0.14	-0.08	-0.14	-0.21	-0.08	-0.15	0.1	0.14
22. Unique URLs/Day (br)	0.21	0.09	0.11	-0.19	-0.11	-0.16	-0.05	-0.11	0.14	0.12
23. Domains/Day (br)	**0.35**	0.18	-0.19	0.09	0.22	-0.01	0.18	-0.02	0.28	-0.04
24. Click Links via Tracking (br)	0.21	0.03	0.14	-0.05	-0.27	**-0.38**	-0.04	-0.21	0.15	0.16
25. Click Links via Source (br)	0.24	-0.16	0	-0.27	0.01	-0.05	0.01	-0.02	0.07	0.08
26. Active Days (NP)	-0.06	0.22	0.17	0.2	0.04	-0.15	0.15	-0.04	0.01	0.02
27. Total URLs/Day (NP)	0.09	**0.35**	0.21	0.04	-0.02	-0.11	0.11	-0.11	-0.02	-0.01
28. Unique URLs/Day (NP)	**0.3**	**0.37**	0.08	0.01	0.02	-0.04	0.12	-0.09	0.05	-0.14
29. Domains/Day (NP)	0.23	0.2	-0.03	-0.12	0.15	0.03	0.12	-0.03	0.07	-0.02
30. Click Links via Source (NP)	**0.34**	0.13	0.15	0.04	-0.19	-0.06	-0.11	-0.21	-0.22	-0.28
31. Total SW	0.03	0.1	**0.31**	0.26	-0.11	-0.18	-0.05	-0.22	0.02	-0.12
32. Vulnerable SW	-0.15	-0.01	0.12	0.15	0	-0.15	-0.02	0.06	0	0.11
33. AV	0.08	-0.09	-0.02	-0.07	0.12	0.06	0.23	-0.12	-0.09	0.01
34. Days Since Windows Update	-0.03	-0.06	0.14	-0.01	-0.09	-0.02	-0.1	0.08	0.05	-0.07

Correlations > 0.3 are bolded.

	11	12	13	14	15	16	17	18	19	20	21	22
1. Malicious URL (br)	0.13	-0.08	-0.14	-0.27	-0.23	-0.31	-0.07	0.1	-0.08	0.26	0.18	0.21
2. Malicious URL (NP)	0.31	-0.13	-0.12	0	0.03	-0.13	0.02	0.09	-0.01	0.16	0.11	0.09
3. Malware	0.01	0.17	0.09	-0.05	-0.06	0.07	-0.02	-0.09	0.32	0.25	0.14	0.11
4. Malicious Files	-0.18	-0.01	-0.17	0.07	0.11	-0.05	0.01	0.06	0.21	0.04	-0.08	-0.19
5. Detection d’	-0.02	-0.15	-0.08	0.02	-0.11	0.02	0.1	0.13	-0.13	-0.03	-0.14	-0.11
6. Detection c	0.15	-0.16	0.21	-0.2	-0.12	-0.13	-0.07	-0.18	-0.15	-0.29	-0.21	-0.16
7. Behavior d’	0.14	-0.21	-0.19	0.02	-0.02	-0.14	0.13	0.09	-0.12	0.11	-0.08	-0.05
8. Behavior c	0.19	-0.16	0.02	-0.26	0.03	-0.25	-0.28	-0.25	-0.07	-0.18	-0.15	-0.11
9. Confidence	0.05	0.11	0.07	0.11	-0.01	-0.11	0.16	0.3	0.07	0.06	0.1	0.14
10. Perceived Consequences	-0.2	0.15	-0.02	0.34	0.24	0.2	0.18	0.15	0.02	0.07	0.14	0.12
11. Male	1	-0.12	0.02	0.19	0.21	-0.02	0.01	0.23	-0.04	0.01	-0.13	-0.09
12. Age	-0.12	1	0.37	0.02	-0.33	0.14	0.27	0.22	0.46	0.33	0.08	0.14
13. College	0.02	0.37	1	0.02	-0.1	0.09	0.09	0.06	0.09	-0.07	-0.07	-0.01
14. SeBIS	0.19	0.02	0.02	1	0.65	0.59	0.57	0.52	0.03	-0.14	-0.18	-0.19
15. Device Subscale	0.21	-0.33	-0.1	0.65	1	0.11	-0.06	0.07	-0.16	-0.38	-0.12	-0.17
16. Password Subscale	-0.02	0.14	0.09	0.59	0.11	1	0.29	0.12	0.14	0.08	-0.09	-0.12
17. Proactive Awareness Subscale	0.01	0.27	0.09	0.57	-0.06	0.29	1	0.34	0.05	0.08	-0.2	-0.15
18. Update Subscale	0.23	0.22	0.06	0.52	0.07	0.12	0.34	1	0.16	0.14	0.07	0.11
19. Time in Study	-0.04	0.46	0.09	0.03	-0.16	0.14	0.05	0.16	1	0.58	0.14	0.24
20. Active Days (br)	0.01	0.33	-0.07	-0.14	-0.38	0.08	0.08	0.14	0.58	1	0.29	0.32
21. Total URLs/Day (br)	-0.13	0.08	-0.07	-0.18	-0.12	-0.09	-0.2	0.07	0.14	0.29	1	0.96
22. Unique URLs/Day (br)	-0.09	0.14	-0.01	-0.19	-0.17	-0.12	-0.15	0.11	0.24	0.32	0.96	1
23. Domains/Day (br)	0.13	-0.19	-0.06	-0.12	-0.02	-0.38	-0.06	0.2	-0.22	-0.1	0.22	0.22
24. Click Links via Tracking (br)	-0.11	0.29	0	-0.1	-0.13	-0.09	-0.05	0.13	0.14	0.18	0.72	0.7
25. Click Links via Source (br)	-0.22	0.11	0.16	-0.17	-0.16	-0.15	0.01	0	0.07	0	0.47	0.53
26. Active Days (NP)	-0.17	0.17	-0.2	0.02	-0.17	0.13	0.18	0	0.53	0.45	0.06	0.06
27. Total URLs/Day (NP)	0.01	0.1	-0.13	-0.11	-0.16	-0.03	-0.07	0.08	0.09	0.28	0.56	0.46
28. Unique URLs/Day (NP)	0.17	0.13	-0.14	-0.12	-0.23	-0.17	0.03	0.24	0.14	0.29	0.25	0.28
29. Domains/Day (NP)	0.2	-0.02	-0.04	-0.06	-0.06	-0.09	-0.04	0.09	-0.24	-0.05	0.17	0.18
30. Click Links via Source (NP)	0.11	0.12	-0.12	-0.1	-0.17	-0.06	-0.03	0.12	0.21	0.33	0	0
31. Total SW	0.04	0.2	-0.02	-0.09	-0.32	0.06	0.05	0.21	0.47	0.41	0.11	0.14
32. Vulnerable SW	-0.07	-0.03	-0.01	0.05	0.1	0.19	-0.1	-0.14	0.35	0.11	-0.17	-0.14
33. AV	0.01	0.25	0.08	0.05	-0.15	0.07	0.16	0.18	0.08	0.13	-0.07	-0.05
34. Days Since Windows Update	-0.07	0.27	-0.08	-0.12	-0.27	-0.08	0.13	0.07	0.36	0.17	0.15	0.12

	23	24	25	26	27	28	29	30	31	32	33	34
1. Malicious URL (br)	0.35	0.21	0.24	-0.06	0.09	0.3	0.23	0.34	0.03	-0.15	0.08	-0.04
2. Malicious URL (NP)	0.18	0.03	-0.16	0.22	0.35	0.37	0.2	0.13	0.1	-0.01	-0.09	-0.04
3. Malware	-0.19	0.14	0	0.17	0.21	0.08	-0.03	0.15	0.31	0.12	-0.02	0.09
4. Malicious Files	0.09	-0.05	-0.27	0.2	0.04	0.01	-0.12	0.04	0.26	0.15	-0.07	0.35
5. Detection d’	0.22	-0.27	0.01	0.04	-0.02	0.02	0.15	-0.19	-0.11	0	0.12	-0.01
6. Detection c	-0.01	-0.38	-0.05	-0.15	-0.11	-0.04	0.03	-0.06	-0.18	-0.15	0.06	-0.13
7. Behavior d’	0.18	-0.04	0.01	0.15	0.11	0.12	0.12	-0.11	-0.05	-0.02	0.23	-0.06
8. Behavior c	-0.02	-0.21	-0.02	-0.04	-0.11	-0.09	-0.03	-0.21	-0.22	0.06	-0.12	-0.17
9. Confidence	0.28	0.15	0.07	0.01	-0.02	0.05	0.07	-0.22	0.02	0	-0.09	-0.1
10. Perceived Consequences	-0.04	0.16	0.08	0.02	-0.01	-0.14	-0.02	-0.28	-0.12	0.11	0.01	0.07
11. Male	0.13	-0.11	-0.22	-0.17	0.01	0.17	0.2	0.11	0.04	-0.07	0.01	-0.07
12. Age	-0.19	0.29	0.11	0.17	0.1	0.13	-0.02	0.12	0.2	-0.03	0.25	0.27
13. College	-0.06	0	0.16	-0.2	-0.13	-0.14	-0.04	-0.12	-0.02	-0.01	0.08	-0.08
14. SeBIS	-0.12	-0.1	-0.17	0.02	-0.11	-0.12	-0.06	-0.1	-0.09	0.05	0.05	-0.12
15. Device Subscale	-0.02	-0.13	-0.16	-0.17	-0.16	-0.23	-0.06	-0.17	-0.32	0.1	-0.15	-0.27
16. Password Subscale	-0.38	-0.09	-0.15	0.13	-0.03	-0.17	-0.09	-0.06	0.06	0.19	0.07	-0.08
17. Proactive Awareness Subscale	-0.06	-0.05	0.01	0.18	-0.07	0.03	-0.04	-0.03	0.05	-0.1	0.16	0.13
18. Update Subscale	0.2	0.13	0	0	0.08	0.24	0.09	0.12	0.21	-0.14	0.18	0.07
19. Time in Study	-0.22	0.14	0.07	0.53	0.09	0.14	-0.24	0.21	0.47	0.35	0.08	0.36
20. Active Days (br)	-0.1	0.18	0	0.45	0.28	0.29	-0.05	0.33	0.41	0.11	0.13	0.17
21. Total URLs/Day (br)	0.22	0.72	0.47	0.06	0.56	0.25	0.17	0	0.11	-0.17	-0.07	0.15
22. Unique URLs/Day (br)	0.22	0.7	0.53	0.06	0.46	0.28	0.18	0	0.14	-0.14	-0.05	0.12
23. Domains/Day (br)	1	0.1	0.17	-0.25	0.23	0.45	0.62	-0.05	-0.03	-0.09	-0.02	-0.22
24. Click Links via Tracking (br)	0.1	1	0.51	0.05	0.41	0.31	0.1	0.09	0.17	-0.17	-0.07	0.18
25. Click Links via Source (br)	0.17	0.51	1	-0.08	0.1	0.18	0.16	0.02	0.02	-0.21	0.1	-0.02
26. Active Days (NP)	-0.25	0.05	-0.08	1	0.22	0.13	-0.31	0.03	0.26	0.18	0.26	0.32
27. Total URLs/Day (NP)	0.23	0.41	0.1	0.22	1	0.73	0.41	0.23	0.2	-0.11	0.05	0.06
28. Unique URLs/Day (NP)	0.45	0.31	0.18	0.13	0.73	1	0.59	0.49	0.29	-0.11	0.05	-0.05
29. Domains/Day (NP)	0.62	0.1	0.16	-0.31	0.41	0.59	1	-0.01	-0.07	-0.21	-0.08	-0.25
30. Click Links via Source (NP)	-0.05	0.09	0.02	0.03	0.23	0.49	-0.01	1	0.27	-0.06	0.14	0.06
31. Total SW	-0.03	0.17	0.02	0.26	0.2	0.29	-0.07	0.27	1	0.25	0.05	0.33
32. Vulnerable SW	-0.09	-0.17	-0.21	0.18	-0.11	-0.11	-0.21	-0.06	0.25	1	-0.09	0.07
33. AV	-0.02	-0.07	0.1	0.26	0.05	0.05	-0.08	0.14	0.05	-0.09	1	0.22
34. Days Since Windows Update	-0.22	0.18	-0.02	0.32	0.06	-0.05	-0.25	0.06	0.33	0.07	0.22	1

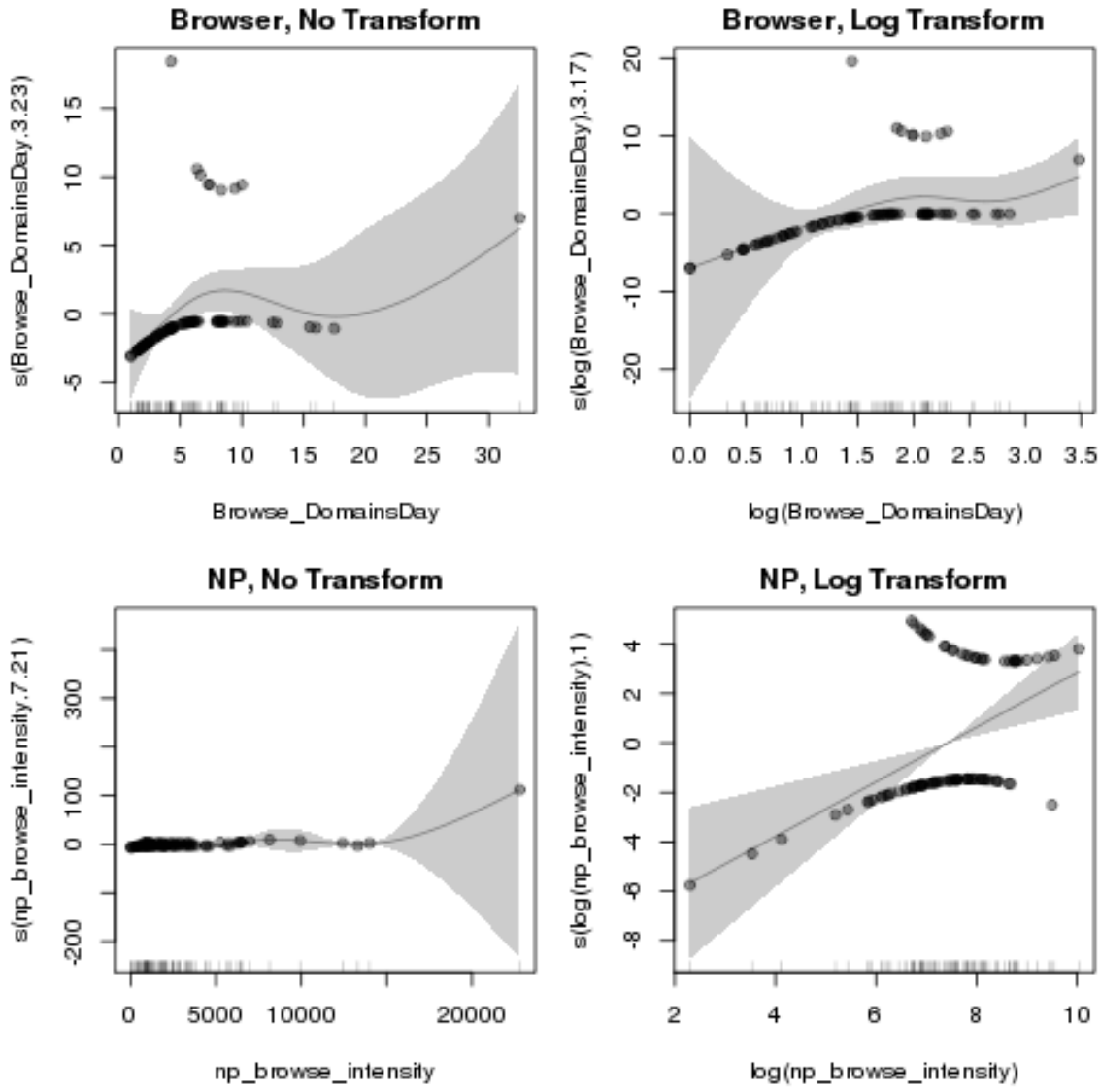

Figure B-3. GAM plot of predictor for browser and network packet data with and without log transformation. In both cases, the log transformation makes the data more linear.

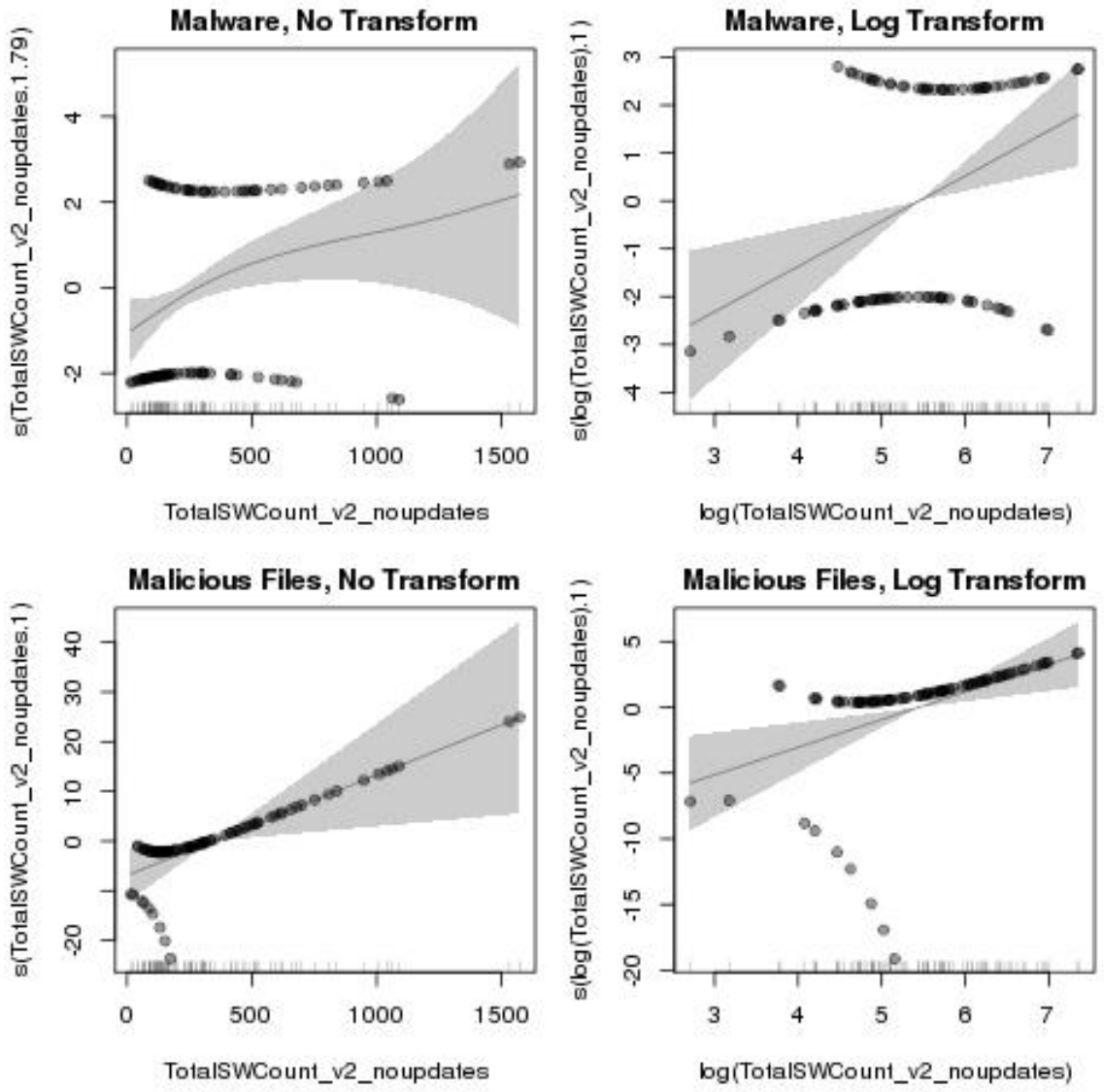

Figure B-4. GAM plot of predictor for malware and malicious file outcomes with and without log transformation. In both cases, the log transformation makes the data more linear.

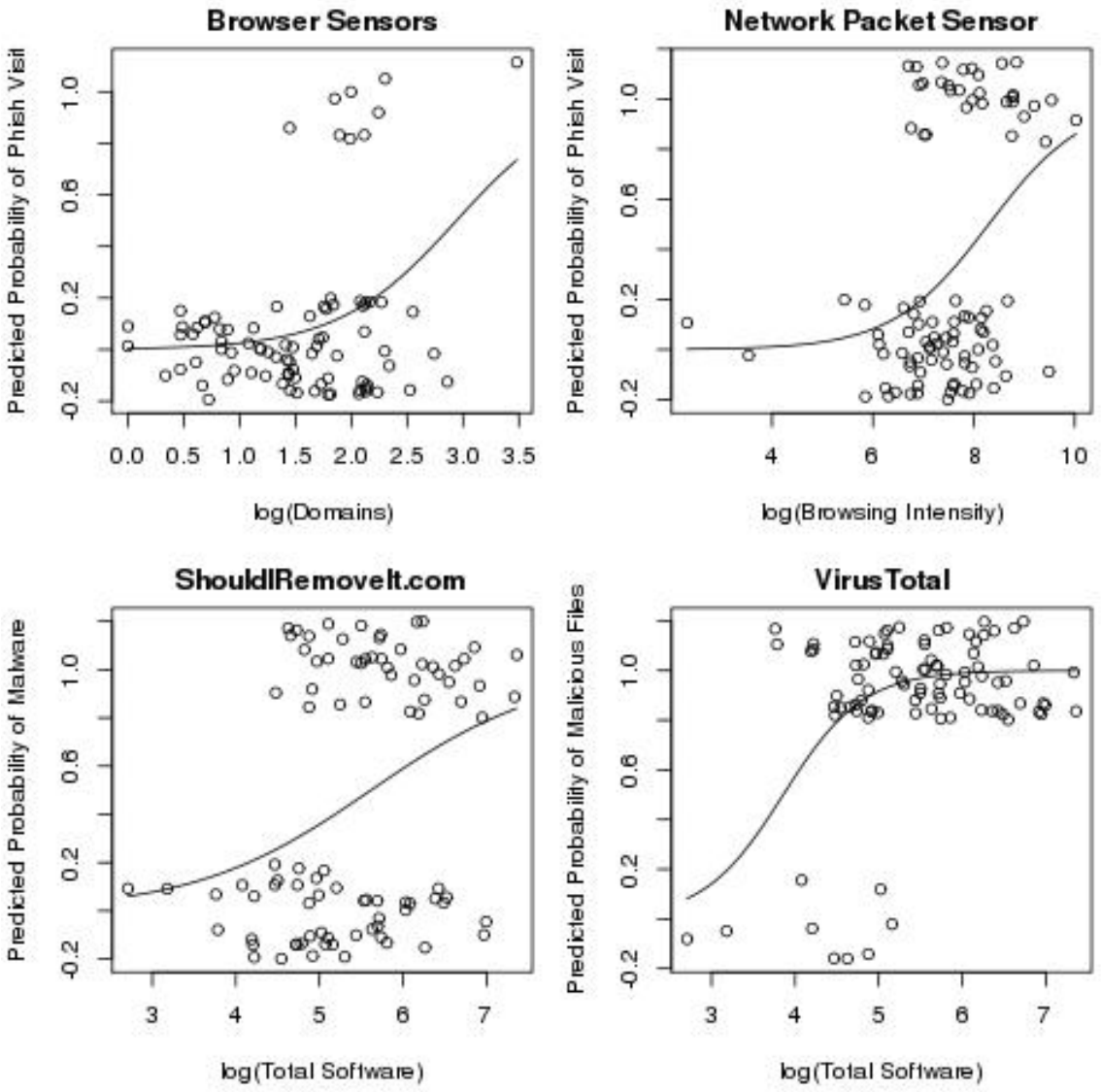

Figure B-5. Plots of each real world outcome with simple regression model (excluding signal detection parameters and demographics).

Table B-14. Logistic regression models and likelihood ratio test (LRT) for each outcome. The predictor was the same as the behavior task models reported in the main text (Tables 4-4,4-5,4-7 and 4-8).

	Malicious URLs (browser)	**Malicious URLs (network packet)**	**Malware**	**Malicious Files**
(Int)	-5.54**	-10.50***	-5.37**	-5.04
	(1.98)	(2.87)	(1.75)	(3.84)
Detection d'	-0.12	-0.29	-0.33	-1.15
	(0.68)	(0.45)	(0.37)	(0.86)
Detection c	-1.04	-0.53	-0.57	-0.80
	(0.84)	(0.56)	(0.58)	(1.35)
Predictor	1.93*	1.42***	0.98**	2.49**
	(0.83)	(0.39)	(0.31)	(0.89)
Age	0	-0.04*	0	-0.05
	(0.03)	(0.02)	(0.02)	(0.03)
Male	0.93	1.55**	0.05	-1.02
	(0.82)	(0.56)	(0.48)	(0.99)
College	-0.69	0.43	0.74	-1.18
	(0.99)	(0.65)	(0.57)	(1.30)
X^2 (LRT)	2.10	1.63	2.41	2.69

C. Chapter 5 Appendix

The R code for this study is available at https://osf.io/2f3yh/.

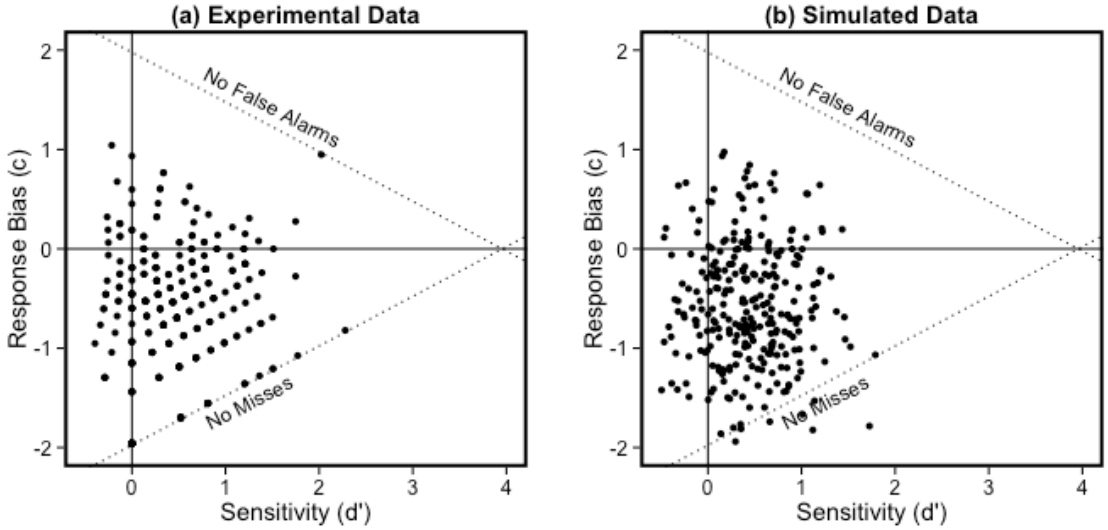

Figure C-1. Validation of simulated sample by comparing it to empirical estimates.

Table C-1. Descriptive statistics for model inputs.

	d'	c	Delta d'	Delta c	Cost of Attack	Cost of FA	Cost of Training
Min	-0.72	-2.11	-2.67	-1.41	18	0	12
Q1	-0.02	-1.17	0.05	-0.36	920	0.26	38
Median	0.30	-0.56	0.57	-0.16	1,800	1	59
Mean	0.33	-0.66	0.57	-0.16	3,000	7.4	60
Q3	0.71	-0.17	1.08	0.03	3,500	3.9	82
Max	1.54	0.75	3.97	1.15	260,000	88,000	110