3-15-2017

Something Looks Phishy Here: Applications of Signal Detection Theory to Cyber-Security Behaviors in the Workplace

Jaclyn Martin
University of South Florida, jmartin85@mail.usf.edu

Something Looks Phishy Here:

Applications of Signal Detection Theory to Cyber-Security Behaviors in the Workplace

by

Jaclyn Martin

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Michael D. Coovert, Ph.D.
Chad Dubé, Ph.D.
Stephen Stark, Ph.D.

Date of Approval:
March 24, 2017

Keywords: cyberpsychology, spear-phishing, phishing, personality, SDT, decision-making

Dedication

This thesis is dedicated to my parents, Suzanne Cox, Ron Martin, and Staci Martin. Without their constant love and support, this thesis would not have been possible. I especially appreciate their encouragement regardless of how many times I called to complain and regardless of the fact that my dad will soon have to concede his place as the highest educated in our family (unless he goes through with that online PhD.). Finally, thank you to Nathan Kowal for listening to many iterations of my thesis, for the unwavering reassurance, and for making the great sacrifice of golfing every Sunday for the past couple years so I could work on my thesis without distraction.

Acknowledgements

I would like to thank Dr. Michael Coovert, who provided invaluable support and feedback over the past two years, in all stages of the thesis project. I greatly appreciate all of his time and efforts throughout the two years of research involved in this project. I would also be remiss if I did not thank Dr. Chad Dubé for answering my endless stream of questions on signal detection theory. Additionally, I would like to thank my committee member, Dr. Stephen Stark for his valuable feedback and direction. A huge thank you to Sarah Frick as well for keeping me on track through our many, many thesis dates. Lastly, I would like to thank my peer (and life) mentor, Rachel Dreibelbis for her constant support and suggestions throughout the thesis process.

Table of Contents

List of Tables

List of Figures

Abstract

Cyber-security is an ever-increasing problem in the 21st century. Though the majority of cyber-security breaches are a direct result of human error (Hu, Dinev, Hart, & Cooke, 2012), there is a dearth of research in psychology on the application of human decision-making for cyber-security compliance. Through an online inbox simulation, the present research examined the utility of a robust psychological model for decision-making, signal detection theory (SDT) for modeling decision-making in the context of receiving and responding to phishing and spear-phishing email scams. The influence of individual differences, specifically conscientiousness, on phishing email detection was also examined. The results indicate that SDT is useful for modeling and measuring cyber-compliance behavior in terms of responding to phishing emails. This finding supports the feasibility of using SDT to monitor training effectiveness for individuals' resistance to social engineering in phishing email detection. There were no significant relationships between participants' scores on conscientiousness and their phishing and spear-phishing email detection ability. Future research should explore predictors of cyber-compliance with regards to individuals' phishing and spear-phishing susceptibility.

Chapter One
Introduction

As Dalal et al. (2010) pointed out, research on the decision-making process is imperative for those in the field of Industrial-Organizational (I-O) Psychology. Decisions drive human behavior in the workplace and are therefore centric to many of the areas of study for I-O researchers. Some decision-making approaches borrow from behavioral economics and examine the influence of human perceptions of risks and rewards associated with certain decision outcomes. Research ranges from modeling the processes through which individuals choose decision-making strategies (Beach & Mitchell, 1978) to evaluating the underpinnings of cognitive shortcuts, or heuristics, used in decision-making (Kahneman, 2011).

One persistent finding in the decision-making literature is that humans do not always act in their best interests (Tversky & Kahneman, 1974). Sometimes their perceptions are skewed (or biased) about what the consequences of a decision might be; in other instances, they might fail to recognize the problems all together. Human performance resulting from these decisions is the subject of many different streams of research across various applications. A recent application of the errors in human decision-making exists in study of cyber-security behaviors (Guo, 2013). That is, human errors in recognizing the cost of certain consequences of their actions related to cyber-security issues (falling prey to social engineering, failing to create strong passwords, etc.) are pervasive. This misjudgment frequently exposes vulnerabilities in organizations' cyber-security, which leaves organizations susceptible to cyber-attacks. These errors are critical as

recent research suggests that over half of successful cyber-attacks are due to human error, not the result of inadequate technology (Hu et al., 2012).

However, the current literature falls short in incorporating *human* factors (e.g. decision-making) in examining common cyber-attacks, such as phishing (El-Din, Cairns, & Clark, 2015). Specifically, there is a lack of empirical and experimental research in this area. We know that employees represent serious threats to cyber-security, but just how susceptible are employees to cyber-threats? Are certain threats easier for employees to detect? Are there individual differences that cause some employees to be more susceptible than others? The present research will address the dearth of research on these human factors relevant to cyber-security vulnerabilities and, additionally, contribute to the overarching research in I-O psychology on cyber-security.

The present research uses a psychological framework to model human error in decision-making in order to better understand cyber-security issues. First, a review of the organizational context of performance in I-O psychology is presented in order to provide a basis for the importance of the present work. Next, current cyber-security issues are discussed in order to build a foundation for the subsequent section on the application of cyber-security issues to decision-making within the I-O framework. Finally, components specific to the present research are detailed and the utilization of Signal Detection Theory (SDT) for modeling cyber-security decisions is presented. Through this lens, employers have the opportunity to take measures for the prevention of employee non-compliance behaviors. Moreover, there exists the potential to impact behavior and increase employee compliance with cyber-security behaviors (Wiederhold, 2014). As such, the application of cyber-security issues to the intersection of decision-making and I-O psychology has direct implications for national security, organizations of all types, and the individuals they employ.

Organizational Context

In I-O psychology, much research is devoted to figuring out how to predict and improve job performance through identifying individual differences that are associated with a difference in job performance (e.g. Dalal, 2005; Schmidt & Hunter, 1998). Though some researchers think of job performance as a unitary concept that broadly encompasses employee behaviors that are relevant to the goals of the organization, research, however, suggests that there are multiple components to job performance and that performance should be considered from a multidimensional perspective. Campbell, Gasser, and Oswald (1996) suggested job performance is a multidimensional construct and split the criterion into declarative knowledge, procedural knowledge, and motivation. Then, Viswesvaran and Ones (2000) found that job performance could be conceptualized as a hierarchy with a general factor at the primary level followed by various dimensions underneath. Currently, researchers conceptualize job performance as having three main components: task performance, counterproductive work behavior, and organizational citizenship behavior. Job performance is separated into these categories on the basis that the behaviors associated with each category either contribute to, or detract from, organizational goals for different reasons (Motowildo, Borman, & Schmit, 1997).

Task performance is composed of what has been traditionally considered job performance. Specifically, task performance consists of the core job behaviors, those that directly contribute to the "technical core" of organizational goals. Often, behaviors associated with task performance are outlined in the occupation's job description. The individual differences that best predict task performance are differences in cognitive ability. Higher scores on tests that measure cognitive ability (e.g. SAT, GRE, etc.) are associated with better task performance (Schmidt & Hunter, 1998). Borman and Motowidlo (1993) introduced

organizational citizenship behavior (OCB), or contextual performance, as another type of performance that is separate from task performance because the associated behaviors do not contribute directly to the technical core of the organization, but rather, they are prosocial behaviors that contribute to the organizational environment. OCBs are also more discretionary than task performance behaviors. Being positive and cooperating with coworkers are rarely behaviors that are considered a job requirement. Still, such prosocial behaviors are associated with general job performance. Individual differences in personality best predict contextual performance or OCBs.

Since for the majority of employees (with the exception of those involved in computer security), cyber-security compliance (i.e., inappropriate response to phishing emails) is not a requirement of their job, it therefore aligns with the OCB dimension of job performance. It would follow that personality differences should predict this contextual performance. Still, there are mixed results from studies looking at the relationship between personality traits and phishing susceptibility (Parrish Jr, Bailey, & Courtney, 2009). The subsequent section will explore the importance of cyber-security compliance within the organizational context.

Current Issues in Cyber-Security

Cyber-security is receiving increased attention from the public. Between the widely publicized cyber-attacks at Sony, JP Morgan, Target, and more recently, the US Government's Office of Personnel Management, it is difficult for the average US citizen to ignore news of current threats to information security. The cost of simply cleaning up and remediating the company's affected servers is astronomical. For instance, a cyber-attack recently cost Sony $171 million and this figure is excluding costs associated with lawsuits, loss of customers, etc. (Rosenbush, 2014). Moreover, it is not only large organizations that are affected by cyber-

attacks. Ransomware is currently a popular cyber-attack that plagues small businesses. A ransomware attack involves hackers denying an organization access to information, files, or even smart appliances (e.g. refrigerators) until a sum of money is paid. Across both large corporations and small organizations, the average cost of a cyber-security breach is $412,000 *per incident* (Major, 2014). Given these increasingly widespread attacks, the federal government is investing billions of dollars in boosting cyber-security defenses (Shalal & Selyukh, 2015).

In addition to the financial concerns, a breach represents a grave crisis for the organization's public relations. CyberAlert (www.cyberalert.com) highlighted that organizations struggle to recuperate from the unfavorable company image resulting from breaches due to the tendency for reporters to mark trends in cyber-security breaches. That is, when news of another breach occurs, reporters echo previously reported breaches as well. Though the public tends to view these instances as a result of inadequate technologies, organizations' vulnerabilities lie not only in their hardware and software defenses, but also in the individual vulnerabilities of the organizations' employees. The typical employee will engage in many behaviors, often unknowingly, that pose cyber-security risks for themselves and their employing organizations. For instance, it is common for employees to engage in behaviors, such as creating simple, easy-to-remember passwords like sequential numbers (e.g. "1234") or, oftentimes "password." Thus, cyber-security negligence poses insider threats that can result in the loss of revenue, reputation, and intellectual property.

There are several advantages to investigating the human side of cyber-security along with the technical aspects. First, as aforementioned, the reality is that the majority of information security breaches result from employee behaviors (Hu et al., 2012). Second, as a consequence of the exponential growth of technology, the information about technical defenses that students

learn will likely be obsolete by the time they graduate and enter the workforce. Echoing this sentiment, Gelernter (2015) stated "University computer science departments are in miserable shape: 10 years behind in a field that changes every 10 minutes." On the other hand, human behavior is much more constant. Thus, the argument can be made that, contrary to technical defenses, defenses construed by an understanding of employee behavior in the context of cyber-security compliance will hold through evolutions of cyber-attacks. Finally, as the subsequent section highlights, it is evident that criminals have not only already recognized the role of human behavior in creating vulnerabilities in an organization's system, but also created attacks centered on exploiting these human vulnerabilities.

The method through which criminals target known characteristics of human behavior and traits in order to gain access to sensitive information is called social engineering (Winkler & Dealy, 1995). Specifically, hackers take advantage of human trust to reveal sensitive information from the users by pretending to be legitimate members of the company, and thus, targeting the trust employees have in their coworkers, bosses, etc. For instance, one social engineering technique, called baiting, involves the infecting of USBs with computer viruses and scattering them across a parking lot. The social engineers preyed on the curiosity of employees in this example, knowing that some of them would insert the devices into their work computers, at which point, the login credentials of such employees would be recorded by a keylogger on the USB. Employees at Secure Network Technologies, Inc. found that when 20 USBs were dispersed in a clients' company parking lot, within just three days, employees had collected and used *fifteen* of these USBs (Stasiukonis, 2006).

Characteristics of social engineering that work to the hackers advantage include: the lack of use of high-end technologies, methods that are very easy for someone with low information

technology (IT) expertise to implement, and the vulnerability of end-users to these tactics. For instance, in the aforementioned baiting example, Stasiukonis (2006) emphasized the "convenience" of the tactic, stating that the technicians "never broke a sweat." Effortlessness paired with a seventy-five percent success rate must make an attractive method for criminals!

The widespread success and use of social engineering accentuates the fact that the firewalls and security measures companies invest in are worthless if employees leave the door wide open for hackers and social engineers to enter. The reason all end-users are vulnerable to social engineering is due to the fact that the tactics utilize known characteristics (e.g., biases) of human nature to ensure success. Thus, if criminals use the psychological manipulation of end-users to gain sensitive information, it follows that tenants of psychology would also be useful in the detection and prevention of insider threat.

Phishing

Phishing email attacks are a common form of social engineering designed to influence recipients to open file attachments or click on embedded links that create vulnerabilities for hackers to exploit in order to gain access to sensitive information (Jagatic, Johnson, Jakobsson, & Menczer, 2007). Like baiting tactics, phishing provides social engineers with a low risk and low effort method that can yield great financial gains (Chandrasekaran, Narayanan, & Upadhyaya, 2006).

Social engineers typically conduct phishing attacks using a three-step process (Chandrasekaran et al., 2006). First, they decide on a business to target and collect the email addresses of potential customers or employees for that business, often using the same techniques that spammers use to collect victims' email addresses. Second, the cyber criminals design and send out the emails scams (considered the "hook" of phishing attacks) in mass quantities to the

collected addresses. These emails either have links embedded in them that will direct the victims to an illegitimate website or they have an attachment that when downloaded will infect the victims' devices. In the final step of the process, the social engineers collect credentials from victims (the "catch" from the phishing attack). When victims click on the embedded links and enter their information on the illegitimate websites, their information is stored for the social engineers' use. Alternatively, the criminals gain access to the victims' devices through a virus transmitted to the end-user's computer through executable email attachments, at which point, criminals can access financial or otherwise sensitive information of the victims that can result in various personal or organizational costs.

A typical phishing method is creating emails that mimic banks, credit card companies, or other business that involve monetary transfer (e.g., eBay). There are several different components (or "lures") involved in phishing attacks that incentivize end-users to expose their sensitive information (see Appendix C for examples). The 2012 US-CERT Security Trends Report lists the most common lure themes as business logistical operations (i.e., packaging and shipping), financials (i.e., tax information), customer complaint claims, and travel (United States Computer Emergency Readiness Team, 2012).

Spear-phishing

Another cyber-security threat involves a tactic called spear-phishing, which is, as its name suggests, a more targeted form of phishing aimed at attacking specific individuals within an organization instead of the organization itself. Through this method, cyber criminals commonly impersonate trusted users and personalize messages for a more sophisticated and elaborate social engineering technique (see Appendix D for examples). Hong (2012) noted an example of spear-phishing in which a military official would be sent an invitation for a general's

retirement party containing malicious links or attachments. In instances such as this, cyber criminals exploit specific information (often obtained through social networking sites) about the targeted end-users in order to gain access to their personal information. Understanding individual differences in employees' decision-making processes regarding these risky behaviors will inform organizations of ways to maximize employee compliance with cyber-security policies and ward off such attacks. Before understanding the nature of the relationship between individual differences and cyber-security compliance, it is essential to define cyber-security compliance as a criterion.

Measuring Cyber-Security Compliance

Though there exist numerous studies on phishing susceptibility, there is a lack of consistency in measurement techniques. Parrish et al. (2009) defined phishing susceptibility along two dimensions: 1) the likelihood that an individual will respond to a phishing email (offer his/her information) and 2) the time it takes for the individual to respond to the email. Susceptibility varies along the time continuum because the greater the amount of time after the phishing email is distributed, the smaller the probability that an individual will fall victim (Jagatic et al., 2007). Janet, Mitchell, Robert, and Bradley (2008) used 12 items detailing "risky behaviors" related to phishing emails to collect survey data on student phishing susceptibility. Vishwanath (2015) used a dichotomous measure of susceptibility, recording whether or not participants clicked a hyperlink in a phishing email that was sent to subjects. Sheng, Holbrook, Kumaraguru, Cranor, and Downs (2010) used a more conservative measure, only recording the number of participants that gave their information in response to a phishing email.

One issue with the measures of susceptibility reviewed above is that none of them take into account expectancy factors. That is, researchers were unable to parse out whether the

susceptibility was due to misestimating the risk associated with falling for phishing scams or simply an individual's ability to detect phishing scams. In the present experiment, signal detection theory (SDT) will be utilized to measure performance in detecting cyber-threats in emails. Signal detection theory includes biases such as expectations of risk in the mathematical model of the response process (Anderson, 2001). Lord (1985) made an argument for the improvement of behavioral measurement in applied psychology through the utilization of SDT, noting the superior operationalization for experimental studies and the availability of improvements for accuracy as the primary benefits. As both a model of decision-making and an analytical tool, the theory represents a strong measure for threat detection performance within the context of this experiment. The subsequent section provides an overview of the SDT model and framework.

Signal Detection Theory

SDT is a commonly employed model for describing decision-making cognition. Particularly, SDT is useful for mapping decision-making in conditions of uncertainty. At its simplest form, SDT measures a decision-maker's ability to discriminate between two stimuli, that is to detect when a signal is present versus when there is no signal present—such as distinguishing between a tone versus background noise (Macmillan & Creelman, 2004). A knowledge of the underpinnings of the model allows researchers to quantify the decision-making process and, moreover, to influence the decision-maker in order to optimize performance. This section provides a brief historical background of SDT, an overview of the current model, and a discussion of current and potential applications of SDT.

Background

SDT has a long, rich history and is sometimes considered one of the sources of cognitive psychology (Macmillan & Creelman, 2004). Radar researchers first introduced the framework for the model in the 1950s within the context of air traffic control (Marcum, 1947; Peterson, Birdsall, & Fox, 1954). Air traffic controllers have to detect, based on radar images on their screen, aircrafts that need to be monitored. The difficulty lies in the copious amounts of "ground clutter" or unwanted echoes in that are returned from things such as animals and atmospheric turbulence that make it more difficult to distinguish echoes that represent the target (an aircraft) versus noise (clutter). Peterson et al. (1954) provided a method for quantifying the ability to distinguish between targets and noise. The researchers outlined four distinct outcomes to any given decision: correct detection, correct rejection, miss, and false alarm. In the case of the air traffic controllers, they could correctly recognize an incoming plane (correct detection), correctly recognize that a flock of birds is not a plane (correct rejection), incorrectly perceive a plane to be a flock of birds (miss), or incorrectly perceive a flock of birds to be a plane (false alarm).

After early radar researchers (Marcum, 1947; Peterson et al., 1954) provided the inception of the idea of quantifying the ability to discern between signal and noise, cognitive researchers (Green & Swets, 1966; Tanner Jr & Swets, 1954) swiftly adapted this framework to fit the context of human decision-making. Much of their research stemmed from threshold theory, which is a probabilistic model of decision-making proposed in the 1800s. Threshold theory proved insufficient for modeling signal detection because its probabilistic nature caused the computed statistics to be variant to changes other than sensitivity (Macmillan & Creelman, 2004). Threshold measures varied within person from trial to trial, even if the same stimulus was used across observations. To address this limitation, Green and Swets (1966) developed the

framework for SDT. This framework is advantageous over threshold theory because the latter proved unsuccessful at separating the influence of the sensitivity and response bias, whereas SDT uses a flexible decision criterion that accommodates variation in components of the model other than sensitivity (described below; Green & Swets, 1966; Macmillan & Creelman, 2004).

Signal Detection Theory Model

Visual representation facilitates the explanation of signal detection theory. As such, this section uses Figure 1 to depict a model of signal detection theory. Two normal distributions represent the probabilities the occurrence of two different events (signal and noise, and noise alone). The criterion (labeled neutral criterion in Figure 1 and commonly denoted by β) represents the cutoff point for the decision. That is, on the right side of the criterion, the person would decide to respond as if the signal were present while on the left side of the criterion, the person would decide to respond as if there were no signal present. The area under the two curves represents the four different possible outcomes (outlined in Table 1) to responding to one of the two events. On the left side of the criterion, where the decision-maker responds as if there is no signal, the individual would either make a correct rejection or miss the signal that was present (miss). Likewise, on the right side of the criterion, where the decision-maker responds as if there is a signal, he or she would either correctly detect the signal (hit) or respond when there was no signal present (false alarm).

Table 1. *The four outcomes of a decision, according to signal detection theory.*

	Respond "Absent"	Respond "Present"
Signal Present	Miss	Hit
Signal Absent	Correct rejection	False Alarm

Criterion. The decision-maker's biases determine the placement of the criterion (the vertical line in Figure 1 labeled Neutral criterion). That is, the criterion might move away from neutral toward the left if the decision-maker believes there is a greater cost to missing a signal than to responding to a signal that is not present. For instance, radar detectors in a warzone would likely perceive there to be a greater cost to missing an enemy plane than to mistaking a flock of birds for a plane. The shift of the criterion to the left represents a liberal bias. Alternatively, the criterion might shift right of the neutral position (a conservative bias) when the decision-maker perceives a greater cost to responding to a signal that is not present.

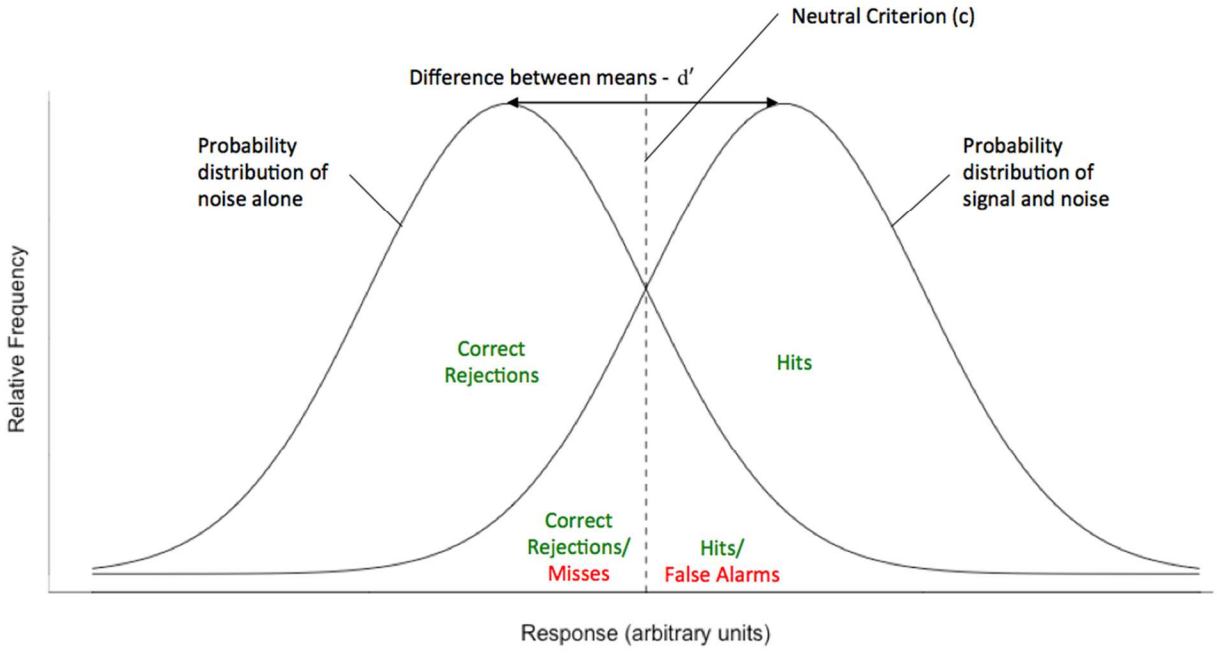


Figure 1. Model of signal detection theory.

D-prime. D-prime (d′) is the most commonly used sensitivity index for SDT (Green & Swets, 1966; Macmillan & Creelman, 2004). d′ represents the sensitivity of the decision-maker

to detection of the signal, where high sensitivity indicates a good ability to distinguish between signal and noise and low sensitivity indicated a poor ability (Macmillan & Creelman, 2004). That is, a high d′ would indicate that a radar detector is more sensitive to detecting the difference between a plane and a flock of birds. Mathematically, d′ measures the distance between the z-score of the hit rate and the z-score of the false alarm rate and, thus, is a measure of the standardized difference between the proportion of hits and the proportion of false alarms (Macmillan & Creelman, 2004). Moreover, d′ can be conceptualized as the differences between the noise and the signal and noise distributions (see Figure 1). This means that the smaller the d′, the more overlap there is between the two events and the more difficult they are to distinguish. Alternatively, with greater distance between the two means (larger d′), it becomes easier for the decision-maker to distinguish between the signal and noise. Consequently, the similarity of the two events affects the size of d′.

ROC curve. Another component of SDT is the Receiver Operating Characteristic (ROC) curve (see Figure 2). ROC curves plot the hit rate that would be obtained for every value of false alarm rate, given a certain d′ value. When d′ is equal to zero, it indicated that performance is at chance level, or the decision-maker has no ability to distinguish between the two signals (this is depicted by the line labeled "d′ = 0" in Figure 2). As the ability for the decision-maker to distinguish between the two signals increases, the curves shift toward the upper left corner, where accuracy is maximized, or the hit rate is always greater than or equal to the false-alarm rate (labeled "d′ = 3" in Figure 2). ROC is typically calculated "by asking participants to supply confidence ratings for their recognition memory decisions" (Wixted, 2007, p. 153).

Payoff and base rate. Two other important components of SDT are payoff and base rate. Payoff is the evaluation of the costs and benefits of each decision (see Table 1). Correct

detections and correct rejections result in benefits to the decision-maker, while false alarms and missed detections result in costs to the decision-maker (Lynn & Barrett, 2014). Base rate is the probability of encountering signals relative to noise (Lynn & Barrett, 2014). As such, payoff and base rate influence the participant's criterion placement or bias.

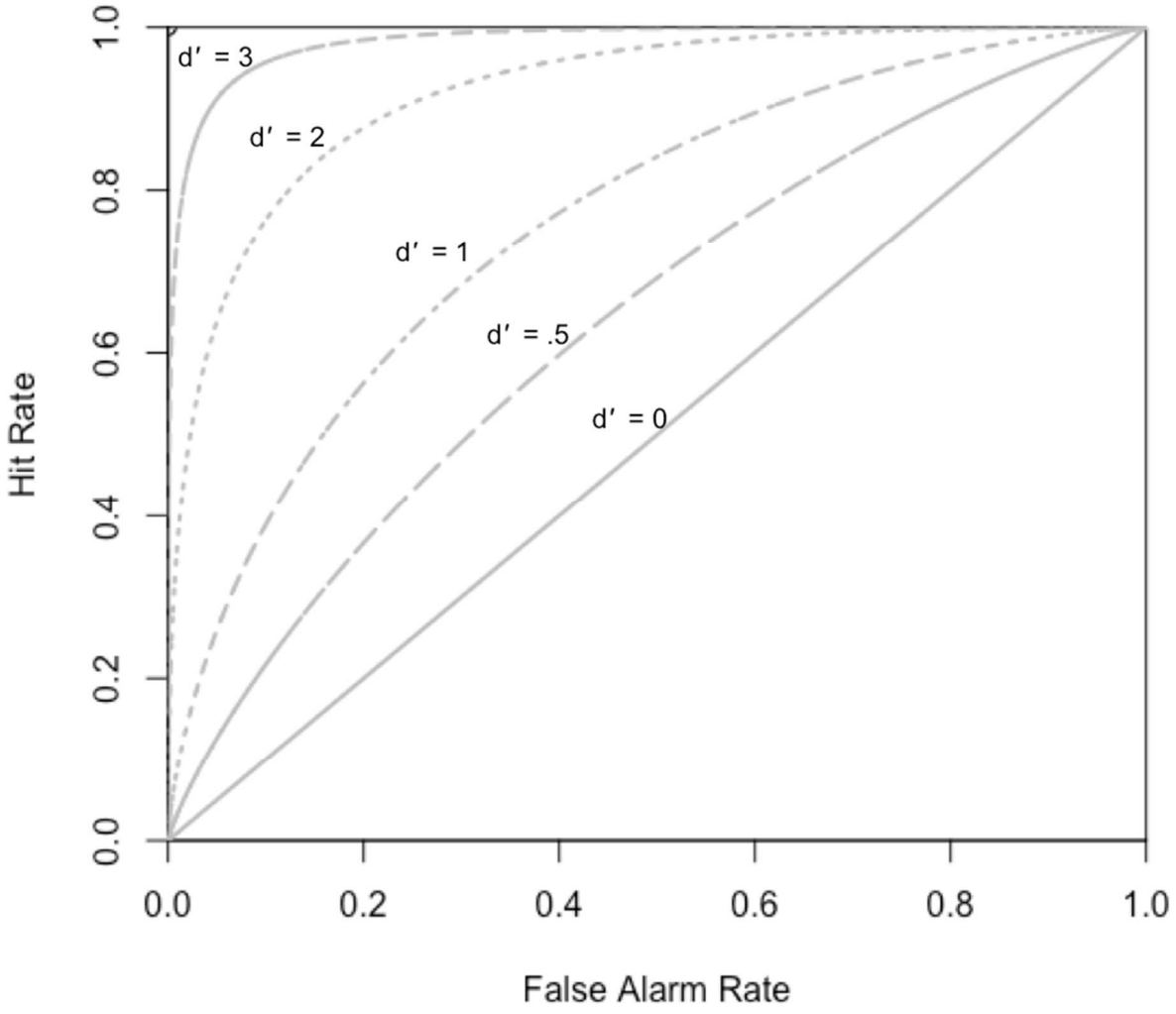


Figure 2. ROCs for SDT on linear coordinates.*

**The closer the curve gets to the upper left corner, the higher the d′ value and the greater the accuracy.*

Important developments in SDT involve studies that considered the manipulation of base rate and payoff. Understanding the manipulation of payoff and base rate is important because it could lead the decision-maker to set the criterion closer the optimal placement where accuracy would be maximized. Healy and Kubovy (1981) conducted an influential study that manipulated each base rate and payoff for participants who were asked to decide whether given values reflected the heights of men or women. The researchers found that base rates had a greater influence on the participants' criterion placement than payoff (Healy & Kubovy, 1981). Another important contribution was the support for the independence of the influences of base rate and payoff information on criterion placement (Bohil & Maddox, 2001; Stevenson, Busemeyer, & Naylor, 1991).

It is worth mentioning that, while the normal-distribution model of SDT discussed thus far is the most commonly used; there are other ways to model SDT (Macmillan & Creelman, 2004). Namely, two other ways are threshold theory, discussed in the background section, and one-interval experiments, which broadens the detection framework to include situations that involve more than two different stimuli or more than two different response types (Macmillan & Creelman, 2004).

Applications of Signal Detection Theory

Traditional applications of signal detection theory involve the radar operators' application (Marcum, 1947), which sparked the development of the theory and cognitive psychology applications (see Green & Swets, 1966; Pazzaglia, Dube, & Rotello, 2013). Currently, SDT is used widely across many fields. For instance, the medical field utilizes SDT to model diagnostics (Lusted, 1971) and alarm fatigue for nurses (Despins, Scott‐Cawiezell, & Rouder, 2010). Another application of SDT is found in detecting management fraud in the field of auditing

(Karim & Siegel, 1998). In the I-O psychology field, Baker and Schuck (1975) used signal detection theory to model rater accuracy in performance assessment for a simulated sales task. Moreover, the researchers found that, through the model, they were able to manipulate rater accuracy in a predictable way by varying base rate and payoff. Similarly, Lievens and Sanchez (2007) used d′ from SDT as a measure of rater accuracy for competency inferences in order to examine then influence of rater training.

Signal Detection Theory and Cyber-Security

Lynn and Barrett's (2014) recent paper advocates for the application of signal detection theory (SDT) through a utility approach that incorporates behavioral economics principles instead of the traditional analytic method. In this way, SDT can be used to predict or explain behavior in decision-making. Lynn and Barrett (2014) mention the applicability of SDT to eyewitnesses' identification of suspects, decisions to place children in foster homes, and cancer detection. In their paper, the authors specifically apply SDT to social-threat detection in evaluating whether a person is angry or not, based on the extent to which their face appears to be scowling. Thus, Lynn and Barrett (2014) advocate for the broadening of application of signal detection theory to wider range of disciplines through a utility approach.

I argue principles from signal detection theory (SDT; Green & Swets, 1966) can measure how susceptible employees are to cyber-threats by disentangling employee decision-making. Thus, I explored potential applications of the SDT framework for measuring and modeling employee cyber-security compliance. SDT is particularly useful for studying decision-making in environments with perceptual uncertainty and risk. Uncertainty occurs when signals are difficult to discriminate from noise. For example, suppose an employee receives an email from the human resources department asking for his or her login information, and the employee is unsure whether

the email is legitimate or if it is a spear-phishing email. Risk is present when incorrectly classifying a signal as noise, and likewise classifying noise as a signal carries some cost. In this case, if the employee decides that it is a legitimate email and responds to the email with information, yet it actually is a phishing email, then the employee has likely cost the company important information and/or monetary resources by granting a hacker access to the organization's database and servers.

In psychophysics, classical (equal-variance) SDT uses two indexes to model decision-making: sensitivity (d′) and response criterion (β; Green & Swets, 1966). Sensitivity (d′) reflects a person's ability to detect signals relative to noise. In the context of cyber-security, sensitivity is an employee's ability to discriminate between threats and non-threats. The response criterion reflects biases that influence a person's tendency to respond to a signal. The placement of the criterion (β) is typically described as liberal, neutral, or conservative. A liberal criterion indicates that a person would be more likely to perceive a threat in an instance where an individual with a conservative criterion placement would perceive noise. That is, if an employee with a liberal criterion receives a borderline suspicious email, he or she would more likely perceive the email as a scam (or believe there is a cyber threat) than perceive the email as safe. This leads to the first research question:

Research Question 1: Do individuals exhibit different sensitivity in detecting spear-phishing and phishing emails?

Though social engineers developed spear-phishing in the hopes of increasing their cyber-attack success rates, there is little empirical research comparing the difference in success rates for spear-phishing versus phishing emails. (Hong, 2012) suggests that end-users are 4.5 times more likely to fall prey to spear-phishing attacks than to phishing attacks. It would make sense

that the use of social networking data to personalize the email and use existing contact information would effectively exploit end-users' trust and cause them to fall prey to the spear-phishing attack.

On the other hand, one could make the argument that the use of private, personal information will create a sort of cocktail party effect (Cherry, 1953), where end-users' will attend to the information in spear-phishing emails more carefully, as the email contains their personal information. In this way, they will be more careful in assessing whether or not the email reflects a threat. Thus, two competing models were proposed:

Hypothesis 1a: *Individuals have a lower sensitivity to detecting a threat in a spear-phishing email than a phishing email.*

Hypothesis 1b: *Individuals have a higher sensitivity to detecting a threat in a spear-phishing email than a phishing email.*

Individual Differences in Conscientiousness

I expected conscientiousness to predict phishing email sensitivity and spear-phishing email sensitivity as it is a strong predictor of organizational citizenship behaviors or extra-task performance that goes beyond core job performance (such as cyber-security compliance; (Chiaburu, Oh, Berry, Li, & Gardner, 2011).

Hypothesis 2. Personality characteristics (specifically the trait of conscientiousness) predict individual's sensitivity to detect phishing and spear-phishing emails, such that those higher on the conscientiousness trait will have a higher sensitivity to detect phishing and spear-phishing emails.

Accompanying an understanding of human decision-making and organizational processes, the field of I-O psychology is in a unique position to address cyber-security issues that

result from social engineering methods. Furthermore, the nature of I-O psychologists' work in handling sensitive employee data presents a need for increased cyber-security knowledge of best practices. The need for research at the intersection of psychology and cyber-security led to recent advances in the literature on the subject (see Chen et al., 2014; Crossler et al., 2013; Steinke et al., 2015; Willison, 2006). The present research extends the growing body of research on cyber-security by addressing some of the critical junctures between these avenues of research through the use of the strong psychological paradigm, SDT.

Chapter Two
Method

Participants and Procedure

The sample consisted of 344 individuals recruited though Amazon's Mechanical Turk (MTurk) system. Although the use of MTurk for organizational research is fairly recent, there is initial support for its use (Landers & Behrend, 2015). Out of the 344 people sampled, 282 cases were included in the analyses because the cases had to meet several criteria for inclusion. First, participants that failed the attention check question (10.5%) were removed. Second, participants that spent less than 10 minutes on the survey (3.5%) were removed. Third, participants with more than 20% of their responses missing for the email task (4.7%) were removed.

Of those participants included in the analyses, 50.2% were female and an average of 35.27 years of age (*SD* = 9.75). Participants worked an average of 38.7 hours per week (*SD* = 9.01). Almost half (43.0%) of participants had their bachelor's degree as their highest form of education, 33.0% had some college, 10.8% had a high school education, 7.5% had vocational school education, 3.9% had a master's degree, 1.1% had a professional or doctoral degree, and 0.7% had less than a high school education. Moreover, 8.5% of participants either did hold or had held a computer security-related job.

First, participants chose the HIT (human intelligence task) from a list on the MTurk website. They then received an anonymous link to the Qualtrics survey. They were prompted to give consent to participate in the study by reading and acknowledging the online consent form. The consent form also informed the participants that their compensation is contingent upon

correctly answering questions an attention check to see if they carefully read and understood instructions. Participants then proceeded to complete the survey, which consisted of an inbox task, conscientiousness items, and demographics.

In order to evaluate the influence of two different types of phishing scams on decision-making, an inbox simulation task was created through Qualtrics. Participants were given a prompt detailing an identity as a human resources assistant to role-play during the experiment (see Appendix E for the prompt) and instructed to go through all the emails in their inbox and respond to the corresponding questions. This experiment employed a 3 (type of email: no threat, spear-phishing, or phishing email) x 2 (response: safe or unsafe) stimulus-response design. The stimuli were presented with base rates that approximate the prevalence rate of non-threatening to threatening emails to the extent that design constraints allowed. That is, 40 emails were created that were non-threatening, as well as 10 emails for each the spear-phishing and phishing categories (see Appendixes C and D for sample stimuli). I adapted these emails from real phishing email scams provided in Cornell University's "Phish Bowl" (http://www.it.cornell.edu/security/phishbowl.cfm). Thus, a series of 60 trials in the form of emails in an inbox were presented to each participant for judgment.

These emails were presented to the participants in batches of 15 emails each. Due to the size of the email images, the survey was only available on a desktop computer (access from mobile devices was blocked through Qualtrics). The emails were randomized within each block to prevent order effects. The participants were not informed which emails were and were not threatening. After reading each email, the participants were prompted with two behavioral response options in which they would indicate how they would respond to the email. One response option listed behaviors that are safe responses to the email (e.g. responding to the

sender in person), while the other response option listed behaviors that are unsafe responses (e.g. click on a URL in the email or reply to the email). Participants then were prompted to estimate their confidence that they would perform the chosen response given the email. Participants next responded to items measuring conscientiousness and demographic questions. Participants were credited $3.00 through Amazon Turk within 48 hours of survey completion. This rate was based off of median rates offered for surveys 45 minutes in length on the MTurk site at the time of data collection. Participants spent an average of 36.09 minutes on the survey (*SD* = 17.03).

Measures

Threat Detection Performance

After reading each email, the participant were prompted to answer the question "How would you handle this email?" with one of two behavioral response options: "Option 1: Reply by email, Download the attached file, AND/OR Click on the selected link in the email (the one that the browser hand is pointing to)" and "Option 2: Delete the email AND/OR Contact the sender by phone or in person." The response information was used to calculate the d′ values for both phishing and spear-phishing emails as a metric for threat sensitivity or threat detection performance. More details on the calculation of this measure are provided in the analysis section below.

Confidence ratings

After answering the threat detection question, the participant was prompted with a slider confidence scale marked with percentages and asked, "On a scale from 0 to 100, how confident are you that this is an appropriate response to the email? (Note: values closer to 0 indicate you are fairly certain it is not an appropriate response and values closer to 100 indicate you are very

certain it is an appropriate response).” This measure was used to compute the group ROCs for analysis, as described below.

Conscientiousness

The personality trait, conscientiousness, was measured using International Personality Item Pool’s (IPIP) 20-item scale (see Appendix A; Goldberg, 1999; Goldberg et al., 2006). The scale had good internal consistency in this study, $\alpha = 0.95$.

Demographics

There were three screening questions asking: 1) whether they have any education in computer security, 2) whether they hold, or have held, a job involving computer security, and 3) whether they have ever helped someone fix a computer problem (adapted from Downs, Holbrook, & Cranor, 2006). If they answered yes to any of these questions, they were asked to explain what they do/had done. Participants whose explanations include an indication of knowledge of computer security, such as ethical hacking, troubleshooting, repairs, or installation of computer systems, etc., were flagged as having security experience. Additionally, gender, age, and ethnicity demographics were collected (see Appendix B).

Chapter Three
Results

Modeling Phishing Email Detection

Hit rates were computed in order to make a simple comparison of participants' performance in threat detection, or threat sensitivity, across the different types of emails (see Table 2). This provided initial insight for the answer to research question 1. The hit rates were higher in the spear-phishing email conditions (40.1%) than phishing email conditions (65.9%), indicating people tend to be more accurate in detecting phishing scams than spear-phishing scams.

Using the SDT model for classification experiment analysis (Macmillan & Creelman, 2004), sensitivity (d′) was computed for each stimulus pair. Assuming unidimensionality of the stimuli—participants only judged the emails by the level of perceived threat—the sensitivity between the remote stimulus pair d'(S_1,S_3) and the sensitivity of the adjacent stimulus pair d'(S_1,S_2) was compared. In other words, the ability for participants to distinguish between no threat and a phishing email and to distinguish between no threat and a spear-phishing email was measured by using the hit rates and false alarm rates to compute each participant's d' (for both spear-phishing and phishing emails) in Excel, as detailed in Table 4 (Macmillan & Creelman, 2004; Sorkin, 1999; Stanislaw & Todorov, 1999).

A paired samples t-test was used as a method for testing the significance of the difference in participants' detection of spear-phishing (M = .42, *SD* = .86) and phishing emails (M = 1.10, *SD* = .93). Each participant's d′ values were used to compare means corresponding to both the

spear-phishing and phishing stimulus pairs ($d'(S_1,S_2)$ and $d'(S_1,S_3)$). The t-test indicated a significant difference in the detection level for the two different types of emails, such that spear-phishing emails were more difficult to detect; $t(281) = 17.97$, $p < .001$. That is, participants have better threat discrimination, or higher sensitivity, with regard to phishing emails than spear-phishing emails, consistent with Hypothesis 1a.

Receiver Operating Curve (ROC) analysis was utilized for further comparison of the sensitivity of spear-phishing versus phishing emails. First, the frequency of each response corresponding to each stimulus was computed (see Table 3). The responses were not only divided into the two response options used for the hit rate, false alarm rate, and d' calculations, but the confidence ratings were used to add additional criteria, which allowed for a closer estimation of the ROCs. The confidence ratings were dichotomized into "sure" and "unsure" response categories based on each participants' average confidence rating. For instance, if a participant's confidence rating for a spear-phishing email stimulus was 45 and their average was 60, then that confidence rating was coded as "unsure". Participants who did not use all the confidence ratings (for all 60 emails) were excluded from this analysis, leaving 207 participants. On average, the participants had 87.8% confidence in their responses ($SD = 10.45$). The frequency data for each stimulus-response pair are presented in Table 3. Then, in accordance with the rating design, the frequency data was used to compute proportions for each stimulus-response pair, transforming the data into hit and false alarm rates (see Table 5). These values were cumulated for use in plotting the ROCs (see Table 6). The hit and false alarm rate pairs were then used to plot the observed data points for the ROCs (see Figure 3).

Maximum-likelihood (ML) estimation was used to generate signal detection theory parameters that best fit the data (Dorfman & Alf, 1969). The parameter estimates for each model

are displayed in Table 7. A chi-square difference test was used between the chi-square statistics produced by the ML estimations for the ROC curve generated from the data and the constrained model where the mean performance (d′) is assumed equal for both spear-phishing and phishing. The difference test showed that there was a significant difference in model fit between the full model (see Figure 3) and the constrained model (see Figure 4); $\chi^2 = 368.35$, $p<.0001$. Thus, both the paired sample t-test and the ROC curve chi-square difference test supported Hypothesis 1a, demonstrating significant individual- and group-level differences in phishing email sensitivity and spear-phishing email sensitivity.

Individual Differences

Correlations were computed for the focal study variables and potential control variables to examine their relationships (see Table 9). Means, standard deviations, skewness, and kurtosis were also computed (see Table 8). Conscientiousness (M = 78.74, *SD* = 15.16) was not significantly related to phishing email sensitivity (M = 1.10, *SD* = .93) or spear-phishing email sensitivity (M = .42, *SD* = .86); r = -.02, $p > .05$, r = -.02, $p > .05$, respectively. Income, gender, and computer security experience were all significantly related to spear-phishing email sensitivity and computer security experience was also significantly related to phishing email sensitivity. Therefore, those variables were entered into the regression equations as controls to further test Hypothesis 2. The regression results (see Table 10) indicated that conscientiousness was not predictive of spear-phishing sensitivity (β = -.03, $p > .05$), even after controlling for income (β =.12, $p <.05$) and computer security experience (β =.13, $p <.05$), which were both significant predictors. Similarly, the regression results indicated that conscientiousness was not predictive of phishing sensitivity (β = -.03, $p > .05$), even after controlling for computer security

experience (β =.20, p <.01), which was a significant predictor. Therefore, Hypothesis 2 was not supported.

Chapter Four
Discussion

The primary purpose of the present research is to examine the applicability of the robust psychological framework, signal detection theory, to certain cyber-security behaviors. ML estimation of SDT parameters showed that SDT parameters fit the observed data well, meaning that SDT parameters are useful for representing and interpreting decision-making in the context of phishing email detection. The SDT measure for sensitivity, d' is used as a metric of spear- and phishing email detection performance. Consistent with prior research, both individual and group analyses of the difference between spear-phishing detection performance and phishing detection performance indicate that individuals are significantly more accurate at detecting phishing emails than spear-phishing emails (c.f. Hong, 2012). This is intuitive because spear-phishing emails are designed to target certain individuals by being more customized and thereby more persuasive. The ability for SDT measures and parameters to capture this difference in performance denotes the utility of SDT to model individual decision-making in the realm of cyber-security compliance.

In an examination of the role of personality, correlation and regression analyses reveal that conscientiousness is not significantly related to, or predictive of, phishing or spear-phishing email detection. It is, however, possible that range restriction attenuated the relationship between conscientiousness and phishing and spear-phishing email detection because very few individuals scored low on conscientiousness. This could be due to the attention check that filtered out participants based on proper reading of the instructions. Individuals who are less detail-oriented

likely failed the attention check and subsequently were removed from analyses. Still, this finding is inconsistent with the few studies that have examined this relationship (e.g., Hu et al., 2012; McBride, Carter, & Warkentin, 2012). For example, Hu et al.'s (2012) research showed that with survey responses from a sample of 148 MIS and MBA students, dutifulness—a facet of conscientiousness—was related to intention to comply with cyber-security behaviors. McBride et al. (2012), in a sample of 46 MBA students, similarly demonstrated that conscientiousness significantly and negatively predicts intention to violate cyber-security policies. Perhaps sample differences account for the difference in effect observed between these studies and the present research as only a small percentage of the sample in this study indicated they had a master's level degree (3.9%). Also, the cited studies utilized business students whereas my sample was more heterogeneous.

Moreover, Hu et al. (2012) and McBride et al. (2012) were looking more broadly at cyber-security compliance and deviance behaviors while the present study was concerned specifically with phishing emails. Perhaps cyber compliance and deviance are a different type of OCB performance, while phishing and spear-phishing detection not represent a type of OCB, and therefore are not predicted by conscientiousness. In fact, a recent study (Darwish, El Zarka, & Aloul, 2012) similarly revealed a null effect between conscientiousness and phishing susceptibility for 100 undergraduate students. Instead, they found that neuroticism was significantly related to phishing susceptibility. Darwish et al. (2012) suggested that this relationship might exist because of the relationship between individuals who score high on neuroticism tend to have a decreased ability to detect lies. Darwish and colleagues (2012) also found a significant relationship between openness to experience and sharing personal information on Facebook. This tendency to share personal information may translate into

entering credentials on a site found through a phishing email. Thus, there are conceivably other personality traits that are more important for predicting phishing email susceptibility.

In addition to exploring these predictive relationships, future research should investigate the nature of cyber-security compliance as it relates to phishing and spear-phishing email detection. Further research into the nature of these compliance behaviors could provide insight into the potential predictors of increased detection performance. For instance, if threat detection performance is more closely related to task performance, research suggests that cognitive ability may predict detection performance (Motowildo et al., 1997).

Implications

The primary implication of the ability to utilize SDT to model phishing detection is the utility of the model for exploring the difference between how individuals actually perform with regard to cyber-security compliance versus how they *should* perform. By calculating the payoffs for difference decisions (according to the four stimulus response-pairs), organizations can pinpoint an optimal criterion level. For instance, as aforementioned, cyber-security breaches resulting from incorrectly responding to a phishing email (a miss in SDT) can be very costly to organizations ($412,000 on average; Major, 2014) in terms of information loss, negative public relations, and so forth. On the other hand, if an employee was very concerned about phishing emails and responded cautiously (e.g. calling the sender in person) to every email, this would undoubtedly hinder productivity. Furthermore, there are implications for the individual's *perception* of cost. Specifically, the weightless, abstract nature of data loss through security breaches may cause employees to underestimate the cost of risky cyber-security behaviors. For instance, when evaluating the cost of changing a password an employee might contemplate the cost associated with the considerable time and effort needed to create and memorize a new

password while largely ignoring the potential cost to the organization if inaction results in a data breach.

Thus, it is important to compute the optimum criterion placement to minimize the costs most important to the organization. Depending on which of the various costs (e.g., productivity, possibility of cyber-breach, etc.) that the organization is interested in minimizing, there will be a different minimum. Once this is computed, organizations will be able to determine how to achieve that optimal level of performance. Specifically, organizations could approach this by identifying individual characteristics that predict optimal performance either for selection purposes or for identifying individual to take part in training interventions. Moreover, SDT can be used to model improvements due to training interventions. Overall, SDT provides a powerful, proven, empirical tool (Lord, 1985) that offers superior operationalization for phishing detection performance and the ability to monitor improvements in accuracy.

More broadly, SDT has many implications for the field of I-O psychology. Researchers and practitioners alike are interested in the measurement of accuracy. Whether it be accuracy in performance appraisal ratings (Baker & Schuck, 1975) or the accuracy of ratings used for job analysis (Lievens & Sanchez, 2007), or accuracy in detecting cyber-security threats in emails, SDT offers is useful tool for operationalizing accuracy and/or performance through d′. Furthermore, SDT could be applied to other topics relevant to the field of I-O psychology, such as selection and training. Specifically, Stillman and Jackson (2005) use SDT to analyze Task-Based Assessment Center data to evaluate the reliability of checklist ratings. By using SDT, they were able to not only judge the accuracy and reliability, but also demonstrate, through the use of SDT's bias measure, that checklist ratings tend to be lenient (Stillman & Jackson, 2005). This bias estimate is an invaluable aspect of SDT that other strategies of measuring and evaluating

accuracy and performance do not provide. Another potential application for SDT in the I-O field is the use of the "utilized SDT" model (Lynn & Barrett, 2014), which utilizes economic concepts to transform SDT into a predictive tool. This could be applied to the selection field where it would model the perceptual uncertainty surrounding the weighing of scores on different selection methods and the behavioral risk associated with choosing the wrong candidate.

Limitations

Though the present research addressed topics important for the growing interest in elevating cyber-security in today's increasingly connected society, it has several limitations. First, this study relied on self-report for all the measures and participants could have been inaccurate, inattentive, or misleading in their responses. Although an attention check was used, there is no way to determine the extent to which participants could have been inaccurate, or misleading in their responses. For instance, though the prompt asked how participants "would" handle the email stimulus, participants may have responded with what they "should" do, which would not accurately capture their veridical phishing and spear-phishing email detection performance.

Secondly, payoff perceptions in this simulation likely do not mirror those in reality. It is difficult to convey how much cost (behavioral risk) is associated with false alarms (responding to an email as if it were threatening when it actually is not a threat). Specifically, the cost associated with false alarms in the real world is principally time and productivity; however, in this simulation, participants were likely not as concerned with those variables as they would be if it were interfering with their actual work. Still, if this were the case, it is even more troubling how poorly individuals performed in terms of detecting both the phishing and spear-phishing emails. On average, participants detected about 7 out of the 10 phishing emails and only 4 out of the 10

spear-phishing emails. That makes 9 successful attacks on the average participant. Since it only takes one successful phishing attack to gain access to a system, this is very concerning! Furthermore, the sample was MTurk workers who use HITs (human intelligence tasks), such as surveys as a source of income; therefore, they may have been as concerned about time and productivity as employees answering emails outside of a simulation setting.

Thirdly, though the base rate for threatening emails was lower than that of non-threatening emails, it was likely much higher than the base rate for threatening emails in the typical organization. For an analogue comparison of the impact of this disparity, research on bomb screening at the airport shows that a low base rate can result in what is called the prevalence effect and can greatly impact a person's bias in decision-making (Wolfe et al., 2007). Future research should examine the influence of different base rates on phishing email detection.

Finally, future research should explore the development of interventions in the form of training programs for employees to increase their detection and appropriate response to cyber-security threats. Specifically, the development of incentive programs that utilize SDT principles could reduce the incidence of cyber-security breaches. Social engineering is a pervasive threat to organizations, regardless of the development of new methods of cyber-attacks and cyber-security defenses. Oftentimes, social engineering is the vehicle used to exploit vulnerabilities in a system. Organizations must take the initiative to train individuals to be resistant to social engineering tactics in order to properly defend themselves from costly attacks. More broadly, as a field specializing in human decision-making and organizational efficiency, addressing the scarcity of cyber-security research in the I-O Psychology literature is imperative because of the direct implications for national security, organizations of all types, and the individuals they employ.

Conclusion

The present study investigated the utility of signal detection theory as a method for measuring and interpreting differences in phishing email susceptibility and spear-phishing email susceptibility. A simulated inbox was used to measure phishing and spear-phishing susceptibility within the SDT framework. Results indicated that SDT was a useful framework for modeling decision-making regarding phishing and spear-phishing emails. Additionally, the present research examined the relationship between conscientiousness and phishing spear-phishing email susceptibility. Results indicated that conscientiousness is not a significant predictor of phishing susceptibility. Future research should further investigate individual differences that predict phishing susceptibility and investigate the utility of SDT for evaluating training effectiveness on cyber-compliance behaviors.

References

Anderson, N. H. (2001). *Empirical direction in design and analysis*: Psychology Press.

Baker, E. M., & Schuck, J. R. (1975). Theoretical note: Use of signal detection theory to clarify problems of evaluating performance in industry. *Organizational Behavior and Human Performance, 13*(3), 307-317.

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of management review, 3*(3), 439-449.

Bohil, C. J., & Maddox, W. T. (2001). Category discriminability, base-rate, and payoff effects in perceptual categorization. *Perception & psychophysics, 63*(2), 361-376.

Borman, W. C., & Motowidlo, S. M. (1993). Expanding the criterion domain to include elements of contextual performance. *Personnel Selection in Organizations; San Francisco: Jossey-Bass*, 71.

Campbell, Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of job performance variability. *Individual differences and behavior in organizations*, 258-299.

Chandrasekaran, M., Narayanan, K., & Upadhyaya, S. (2006). *Phishing email detection based on structural properties.* Paper presented at the NYS Cyber Security Conference.

Chen, T. R., Shore, D. B., Zaccaro, S. J., Dalal, R. S., Tetrick, L. E., & Gorab, A. K. (2014). An Organizational Psychology Perspective to Examining Computer Security Incident Response Teams. *IEEE Security & Privacy*(5), 61-67.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America, 25*(5), 975-979.

Chiaburu, D. S., Oh, I.-S., Berry, C. M., Li, N., & Gardner, R. G. (2011). The five-factor model of personality traits and organizational citizenship behaviors: a meta-analysis. *Journal of Applied Psychology, 96*(6), 1140.

Crossler, R. E., Johnston, A. C., Lowry, P. B., Hu, Q., Warkentin, M., & Baskerville, R. (2013). Future directions for behavioral information security research. *computers & security, 32*, 90-101.

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90*(6), 1241.

Dalal, R. S., Bonaccio, S., Highhouse, S., Ilgen, D. R., Mohammed, S., & Slaughter, J. E. (2010). What If Industrial–Organizational Psychology Decided to Take Workplace Decisions Seriously? *Industrial and Organizational Psychology, 3*(4), 386-405.

Darwish, A., El Zarka, A., & Aloul, F. (2012). *Towards understanding phishing victims' profile.* Paper presented at the Computer Systems and Industrial Informatics (ICCSII), 2012 International Conference on.

Despins, L. A., Scott‐Cawiezell, J., & Rouder, J. N. (2010). Detection of patient risk by nurses: a theoretical framework. *Journal of advanced nursing, 66*(2), 465-474.

Dorfman, D. D., & Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating-method data. *Journal of mathematical psychology, 6*(3), 487-496.

El-Din, R. S., Cairns, P., & Clark, J. (2015). The Human Factor in Mobile Phishing. *New Threats and Countermeasures in Digital Crime and Cyber Terrorism*, 53.

Gelernter, D. (2015). Why I'm Not Looking to Hire Computer-Science Majors. Retrieved from http://www.wsj.com/articles/why-im-not-looking-to-hire-computer-science-majors-1440804753

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

Guo, K. H. (2013). Security-related behavior in using information systems in the workplace: A review and synthesis. *computers & security, 32*, 242-251.

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory, 7*(5), 344.

Hong, J. (2012). The state of phishing attacks. *Communications of the ACM, 55*(1), 74-81.

Hu, Q., Dinev, T., Hart, P., & Cooke, D. (2012). Managing Employee Compliance with Information Security Policies: The Critical Role of Top Management and Organizational Culture*. *Decision Sciences, 43*(4), 615-660.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM, 50*(10), 94-100.

Janet, L., Mitchell, D. B. A., Robert, B., & Bradley, K. (2008). Analysis of Student Vulnerabilities to Phishing. *AMCIS 2008 Proceedings*, 271.

Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.

Karim, K. E., & Siegel, P. H. (1998). A signal detection theory approach to analyzing the efficiency and effectiveness of auditing to detect management fraud. *Managerial Auditing Journal, 13*(6), 367-375.

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, mechanical turk, and other convenience samples. *Industrial and Organizational Psychology*, 1-23.

Lievens, F., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *Journal of Applied Psychology, 92*(3), 812.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70*(1), 66.

Lusted, L. B. (1971). Signal detectability and medical decision-making. *Science, 171*(3977), 1217-1219.

Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" Signal Detection Theory. *Psychological science, 25*(9), 1663-1673.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*: Psychology press.

Major, A. (2014). Cyber-Breach & NISPOM Conforming Change 2 – It's What's on the Inside That Counts.

Marcum, J. I. (1947). A statistical theory of target detection by pulsed radar.

McBride, M., Carter, L., & Warkentin, M. (2012). *Exploring the Role of Individual Employee Characteristics and Personality on Employee Compliance with Cybersecurity Policies*. Retrieved from

Motowildo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*(2), 71-83.

Parrish Jr, J. L., Bailey, J. L., & Courtney, J. F. (2009). A Personality Based Model for Determining Susceptibility to Phishing Attacks. *Little Rock: University of Arkansas*.

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Information Theory, Transactions of the IRE Professional Group on, 4*(4), 171-212.

Rosenbush, S. (2014). The Morning Download: Sony Breach Could Cost $100 million. Retrieved from http://blogs.wsj.com/cio/2014/12/10/the-morning-download-sony-breach-could-cost-100-million/

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262.

Shalal, A., & Selyukh, A. (2015). Obama seeks $14 billion to boost U.S. cybersecurity defenses. *Technology*. Retrieved from http://www.reuters.com/article/2015/02/02/us-usa-budget-cybersecurity-idUSKBN0L61WQ20150202

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010). *Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Sorkin, R. D. (1999). Spreadsheet signal detection. *Behavior Research Methods, Instruments, & Computers, 31*(1), 46-54.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149.

Stasiukonis, S. (2006). Social Engineering, the USB Way. *Dark Reading*. Retrieved from http://web.archive.org/web/20060713134051/http://www.darkreading.com/document.asp?doc_id=95556&WT.svl=column1_1

Steinke, J., Bolunmez, B., Fletcher, L., Wang, V., Tomassetti, A. J., Repchick, K. M., . . . Tetrick, L. E. (2015). Improving Cybersecurity Incident Response Team Effectiveness Using Teams-Based Research. *Security & Privacy, IEEE, 13*(4), 20-29.

Stevenson, M. K., Busemeyer, J. R., & Naylor, J. C. (1991). Judgment and decision-making theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, pp. 283-374). Palo Alto, CA: Consulting Psychologists Press.

Stillman, J. A., & Jackson, D. J. (2005). A detection theory approach to the evaluation of assessors in assessment centres. *Journal of Occupational and organizational Psychology, 78*(4), 581-594.

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological review, 61*(6), 401.

Team, U. S. C. E. R. (2012). *US-CERT Security Trends Report: 2012 in Retrospect*.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

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.

Viswesvaran, C., & Ones, D. S. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment, 8*(4), 216-226.

Wiederhold, B. K. (2014). The role of psychology in enhancing cybersecurity. *Cyberpsychology, Behavior, and Social Networking, 17*(3), 131-132.

Willison, R. (2006). Understanding the perpetration of employee computer crime in the organisational context. *Information and organization, 16*(4), 304-324.

Winkler, I. S., & Dealy, B. (1995). *Information Security Technology? Don't Rely on It. A Case Study in Social Engineering.* Paper presented at the USENIX Security.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological review, 114*(1), 152.

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.

Tables

Table 2. Stimulus-Response Matrix

Stimulus	*Response to prompt ("How would you handle this email?")*	
	Safe Behavior	Unsafe Behavior
S_1 = No threat	False alarms (25.5)	Correct rejections (74.5)
S_2 = Phishing email	Hits (65.9)	Misses (34.1)
S_3 = Spear-phishing email	Hits (40.1)	Misses (59.9)

Note. Values are in percentages and reflect the aggregate of responses across participants.

Table 3. Frequency Stimulus-Response Table

	Response				
	Safe Behavior		Unsafe Behavior		Total
	"Sure"	"Unsure"	"Unsure"	"Sure"	
No Threat	1232	874	1638	4536	8280
Phishing	1030	332	271	437	2070
Spear-phishing	504	325	371	870	2070

Note. N = 207. Participants made a binary detection response indicating "Option 1" for safe behaviors and "Option 2" for unsafe behaviors. This response was immediately followed by a confidence judgment. The participants' confidence judgments were coded into sure and unsure based on whether it fell above or below the participants' mean confidence rating. Participants with less than 60 responses for the confidence questions were excluded. The values in each cell represent the number of responses for the stimulus-response pair.

Table 4. Formulas for Calculation of SDT Statistics for Sample Participant

	A (Labels Only)	*Formula (for Column B; Then copy to all other columns)*	*B (Part. 1)*
1	# hits (S_2)		5
2	# hits (S_3)		3
3	# misses (S_2)		5
4	# misses (S_3)		7
5	# false alarms		8
6	# correct rejections		32
7	H_1 (hit rate for S_2)	=IF(B3>0, B1/(B1+B3),(B1 – 0.5)/(B1+B3))	0.500
8	H_2 (hit rate for S_3)	=IF(B4>0, B2/(B2+B4),(B2 – 0.5)/(B2+B4))	0.700
9	F (false alarm rate)	=IF(B5>0, B5/(B5+B6),0.5/(B5+B6))	0.200
10	$z(H_1)$	=NORMSINV(B7)	0.000
11	$z(H_2)$	=NORMSINV(B8)	0.524
12	z(F)	=NORMSINV(B9)	-0.842
13	$d'(S_1,S_2)$	=B10-B12	0.842
14	$d'(S_1,S_3)$	=B11-B12	1.366
15	$d'(S_2,S_3)$	=B14-B13	0.524

Note. The d' values for different stimulus pairs ($d'(S_1,S_3)$ and $d'(S_2,S_3)$) represent the end user's sensitivity for detecting the corresponding threat from the no threat condition, a higher value indicates higher sensitivity. The d' value in row 15 ($d'(S_2,S_3)$) represents the difference in end-user sensitivity between the spear-phishing and phishing email detection conditions.

Table 5. Proportion of Trials on which the Stimulus Yielded Each Response

	Safe Behavior		Unsafe Behavior		Total
	"Sure"	"Unsure"	"Unsure"	"Sure"	
No Threat	0.15	0.10	0.20	0.55	1.00
Phishing	0.50	0.16	0.13	0.21	1.00
Spear-phishing	0.24	0.16	0.18	0.42	1.00

Note. N = 208. Participants made a binary detection response indicating "Option 1" for safe behaviors and "Option 2" for unsafe behaviors. This response was immediately followed by a confidence judgment. The participants' confidence judgments were coded into sure and unsure based on whether it fell above or below the participants' mean confidence rating. Participants with less than 60 responses for the confidence questions were excluded. The values in each cell represent the number of responses for the stimulus-response pair.

Table 6. Hit and False-Alarm Rates: Cumulative Proportions

	Safe Behavior		Unsafe Behavior	
	"Sure"	"Unsure"	"Unsure"	"Sure"
No Threat	0.15	0.25	0.45	1.00
Phishing	0.50	0.66	0.79	1.00
Spear-phishing	0.24	0.40	0.58	1.00

Note. N = 208. The values in each cell represent the cumulative frequencies of responses for the stimulus-response pair.

Table 7. Chi-Square Difference Test of Model Fit.

	μ	σ	c_1	c_2	c_3	χ^2	df	*p*
Model 1: Full								
Phishing	1.06	1.11	1.05	0.65	0.13	3.29	2	0.19
Spear-phishing	0.36	1.03						
Model 2: Constrained								
Phishing	.74	1.17	1.05	0.65	0.13	371.64	4	<.0001
Spear-phishing	.74	1.17						
Model Comparison: $\Delta \chi^2$						368.35	2	<.0001

Notes. μ = parameter mean estimate, σ = parameter standard deviation estimate, $c_{1\text{-}3}$ = parameter criterion estimates.

Table 8. Means, Standard Deviations, Skewness, and Kurtosis of Study Variables

	M	SD	Skewness	Kurtosis
1. Age	35.27	9.75	1.04	.78
2. Gender	1.50	.50	-.01	-2.01
3. Education	4.22	1.10	-.70	.26
4. Response Duration (in minutes)	36.09	17.30	1.47	3.20
5. Income	4.69	1.97	-.05	-.88
6. Number of hours worked per week	38.70	9.01	-.44	2.46
7. Computer and Technology Skills	31.37	18.94	-.09	-.99
8. Computer Security Experience	.61	.77	1.06	.38
9. Conscientiousness	78.74	15.16	-.56	-.34
10. Phishing Hit Rate	.63	.22	-.57	-.18
11. Spear-phishing Hit Rate	.41	.23	.46	-.70
12. False Alarm Rate	.27	.15	.81	1.06
13. Phishing Sensitivity: $d'(S_1,S_2)$	1.10	.93	-.40	.11
14. Spear-Phishing Sensitivity: $d'(S_1,S_3)$	.42	.86	.23	-.09
15. Difference in sensitivity: $d'(S_2,S_3)$	.68	.64	-.32	-.11

Note. Number of hours worked per week does not include those that are unemployed. Gender is a dichotomous variable.

Table 9. Correlations among Study Variables.

	1	2	3	4	5	6
1. Age	-					
2. Gender	-0.23***	-				
3. Education	-0.04	-0.03	-			
4. Response Duration (in minutes)	0.20***	0.00	-0.06	-		
5. Income	0.05	-0.09	0.21*	-0.06	-	
6. Number of hours worked	0.03	0.10	0.10	-0.07	0.20**	-
7. Computer and Technology Skills	-0.07	0.15**	0.22***	-0.05	0.35***	0.17**
8. Computer Security Experience	-0.04	0.13	0.05	0.05	0.09	0.01
9. Conscientiousness	0.16**	-0.14*	-0.13*	-0.04	0.07	0.05
10. Phishing Hit Rate	0.06	0.12*	-0.04	0.02	0.06	0.15*
11. Spear-phishing Hit Rate	-0.04	0.16**	-0.02	-0.02	0.09	0.15*
12. False Alarm Rate	-0.05	0.04	-0.09	-0.06	-0.07	0.02
13. Phishing Sensitivity: $d'(S_1,S_2)$	0.06	0.08	0.03	0.05	0.08	0.11
14. Spear-Phishing Sensitivity: $d'(S_1,S_3)$	-0.02	0.12*	0.04	0.01	0.12*	0.11
15. Difference in sensitivity: $d'(S_2,S_3)$	0.10	-0.04	-0.01	0.05	-0.05	0.01

Notes. N = 234-282. *p<.05, **p<.01, ***p<.001

Table 9, continued. Correlations among Study Variables

	7	8	9	10	11	12	13	14	15
1. Age									
2. Gender									
3. Education									
4. Response Duration (in minutes)									
5. Income									
6. Number of hours worked									
7. Computer and Technology Skills	-								
8. Computer Security Experience	0.29**	-							
9. Conscientiousness	-0.02	0.07	-						
10. Phishing Hit Rate	0.04	0.17*	-0.04	-					
11. Spear-phishing Hit Rate	0.02	0.07	-0.05	0.56***	-				
12. False Alarm Rate	-0.06	-0.13*	-0.02	-0.20*	0.02	-			
13. Phishing Detection	0.06	0.19**	-0.02	0.85***	0.43***	-0.67***	-		
14. Spear-Phishing Detection	0.04	0.14*	-0.02	0.58***	0.80***	-0.56***	0.75***	-	
15. Difference in detection ability	0.04	0.09	0.00	0.45***	-0.46***	-0.23***	0.44***	-0.26***	-

Notes. N = 234-282. *p<.05, **p<.01, ***p<.001

Table 10. Regression Results

	Outcome	
Model	Phishing Detection	Spear-Phishing Detection
Step 1: Controls		
Gender	.06	.11
Income	.07	.12*
Computer Security Experience	.20**	.13*
Step 1 R^2	.05	.05
Step 2: Direct Effects		
Gender	.06	.11
Income	.07	.12*
Computer Security Experience	.20**	.13*
Conscientiousness	-.03	-.03
Total F	3.636**	3.205*
Total R^2	.05	.05
ΔR^2	.00	.00

Notes. N = 234-282. *p<.05, **p<.01. Standardized estimates.

Figures

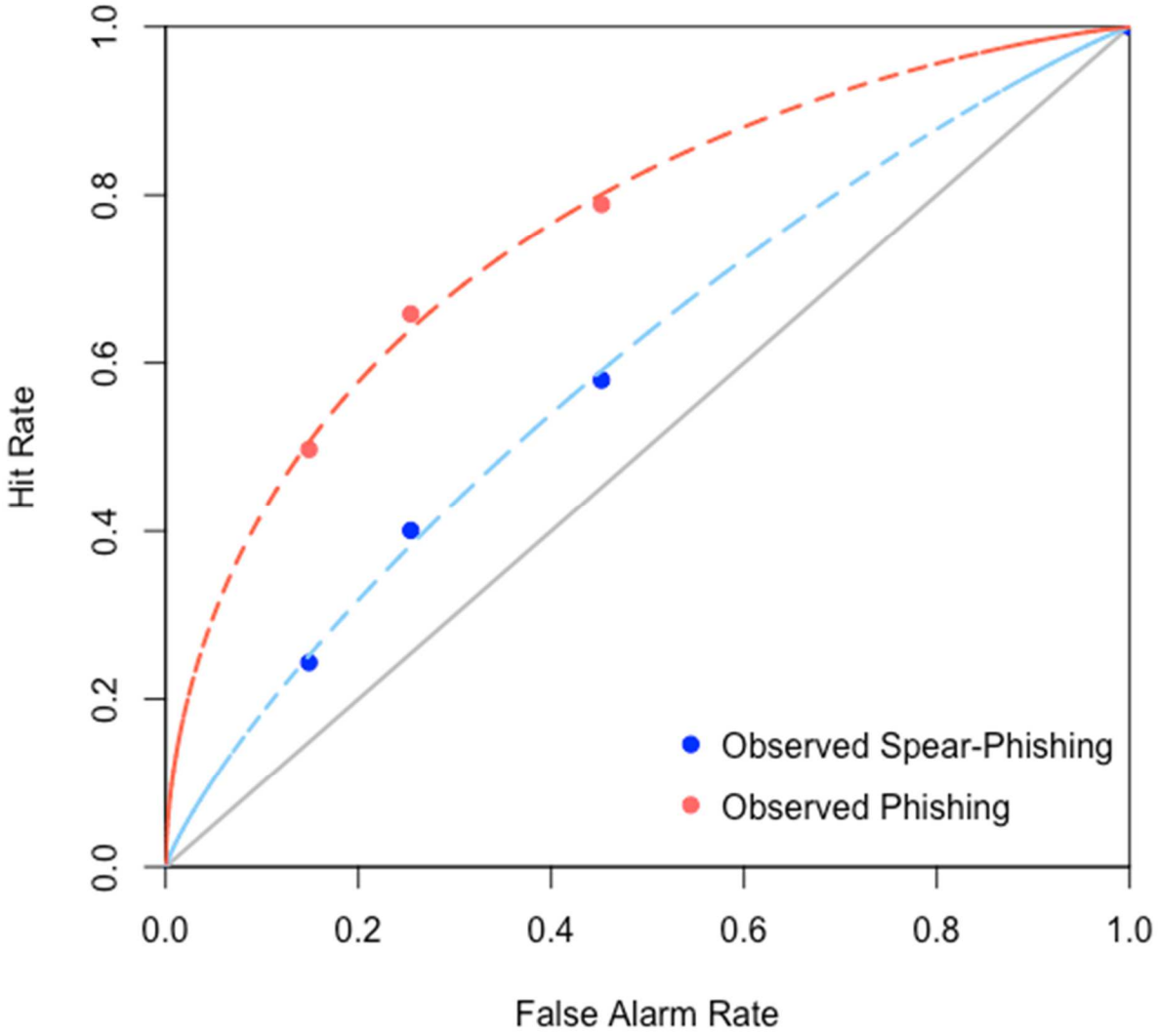


Figure 3. ROCs for Group-level Phishing and Spear-phishing Email Sensitivity Prediction, the full model.

Note. This graph plots the false alarm rates for the non-threatening email stimuli and the hit rates for each the spear-phishing and phishing email stimuli. The blue dots indicate the observed spear-phishing email hit rate and false alarm points while the red dots indicate the observed phishing email hit rate and false alarm points. The dotted lines indicate the fitted ROC curve for the best-fitting signal detection theory parameters. The spear-phishing ROC is also closer to the chance line, indicating that participants were more sensitive to phishing emails.

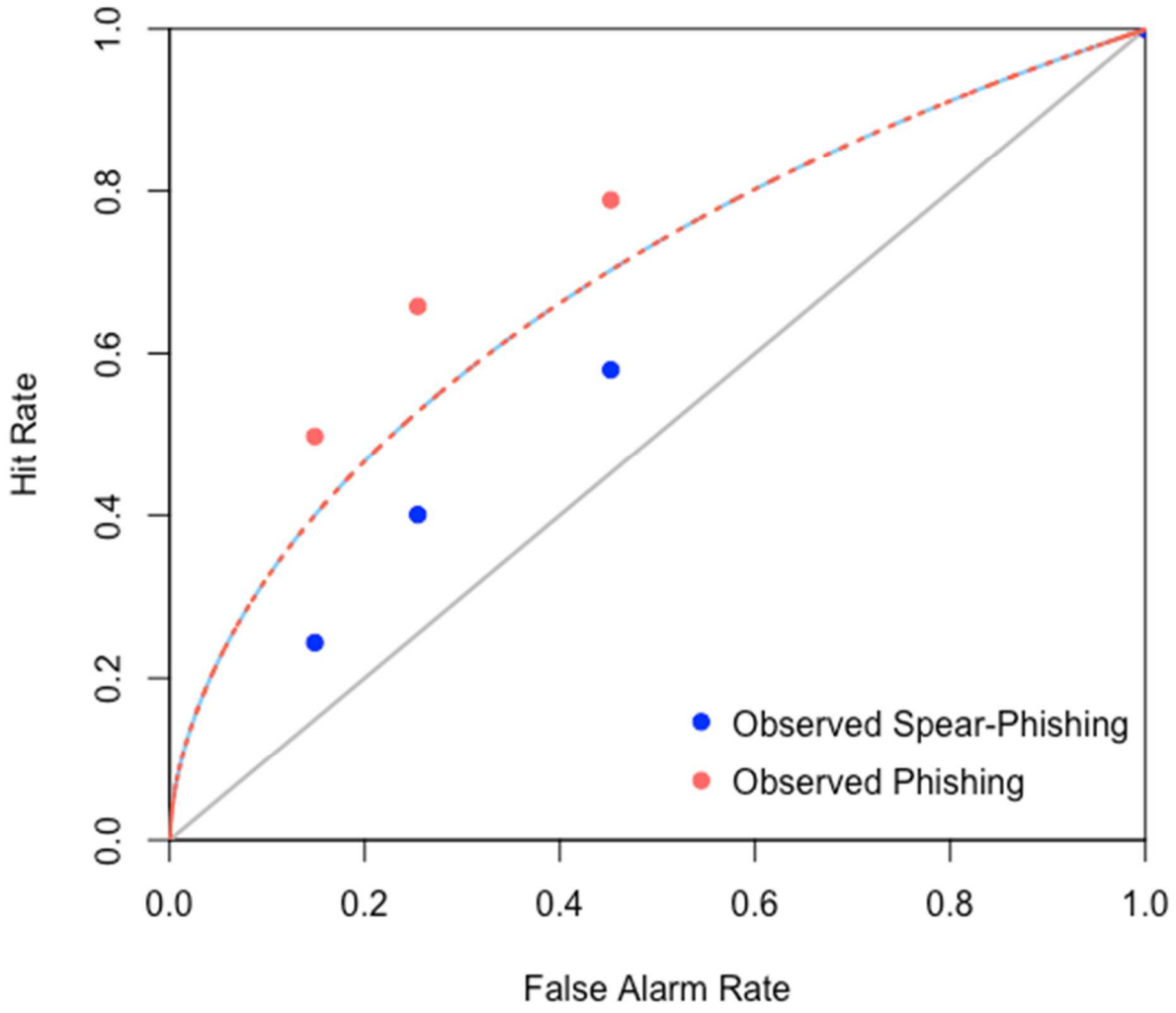


Figure 4. ROCs for Group-level Phishing and Spear-phishing Email Sensitivity Prediction, the constrained model.

Note. This graph plots the false alarm rates for the non-threatening email stimuli and the hit rates for each the spear-phishing and phishing email stimuli. The blue dots indicate the observed phishing email hit rate and false alarm points while the red indicated the observed spear-phishing email hit rate and false alarm points. The dotted lines indicate the fitted ROC curve for the best-fitting signal detection theory parameters. In this constrained model, the means are constrained equal so the parameter estimates are the same. These SDT parameters, where the means are constrained to be equal, do not fit the observed data; $\chi^2(4) = 371.64$, $p<.0001$.

Appendices

Appendix A: Conscientiousness Scale

On the following page, there are phrases describing people's behaviors. Please use the rating scale below to describe how accurately each statement describes *you*. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. So that you can describe yourself in an honest manner, your responses will be kept in absolute confidence. Please read each statement carefully, and then select an option on the scale.

Response Options: 1 = "Very Inaccurate"; 2 = "Moderately Inaccurate"; 3 = "Neither Inaccurate nor Accurate"; 4 = "Moderately Accurate"; 5 = "Very Accurate"

1. Am always prepared.
2. Pay attention to details.
3. Get chores done right away.
4. Carry out my plans.
5. Make plans and stick to them.
6. Complete tasks successfully.
7. Do things according to a plan.
8. Am exacting in my work.
9. Finish what I start.
10. Follow through with my plans.
11. Waste my time.
12. Find it difficult to get down to work.
13. Do just enough work to get by.
14. Don't see things through.
15. Shirk my duties.
16. Mess things up.
17. Leave things unfinished.
18. Don't put my mind on the task at hand.
19. Make a mess of things.
20. Need a push to get started.

Appendix B: Demographic Questions

Please indicate your highest level of education

- ☐ Some high school
- ☐ High school diploma/GED
- ☐ Some college
- ☐ Trade/technical/vocational training
- ☐ Associate's degree
- ☐ Bachelor's degree
- ☐ Master's degree
- ☐ Professional degree
- ☐ Doctoral degree

Do you have any education in computer security?

- ☐ Yes
- ☐ No

If yes, what type of education (certification, bachelor's, etc.)? _____

Do you hold a job involving computer security?

- ☐ Yes
- ☐ No

If yes, what is the job title? _____

Have you held a job involving computer security?

- ☐ Yes
- ☐ No

If yes, what was the job title? _____

Have you ever helped someone fix a computer problem?

- ☐ Yes
- ☐ No

If yes, what assistance did you provide? _____

What is your gender?

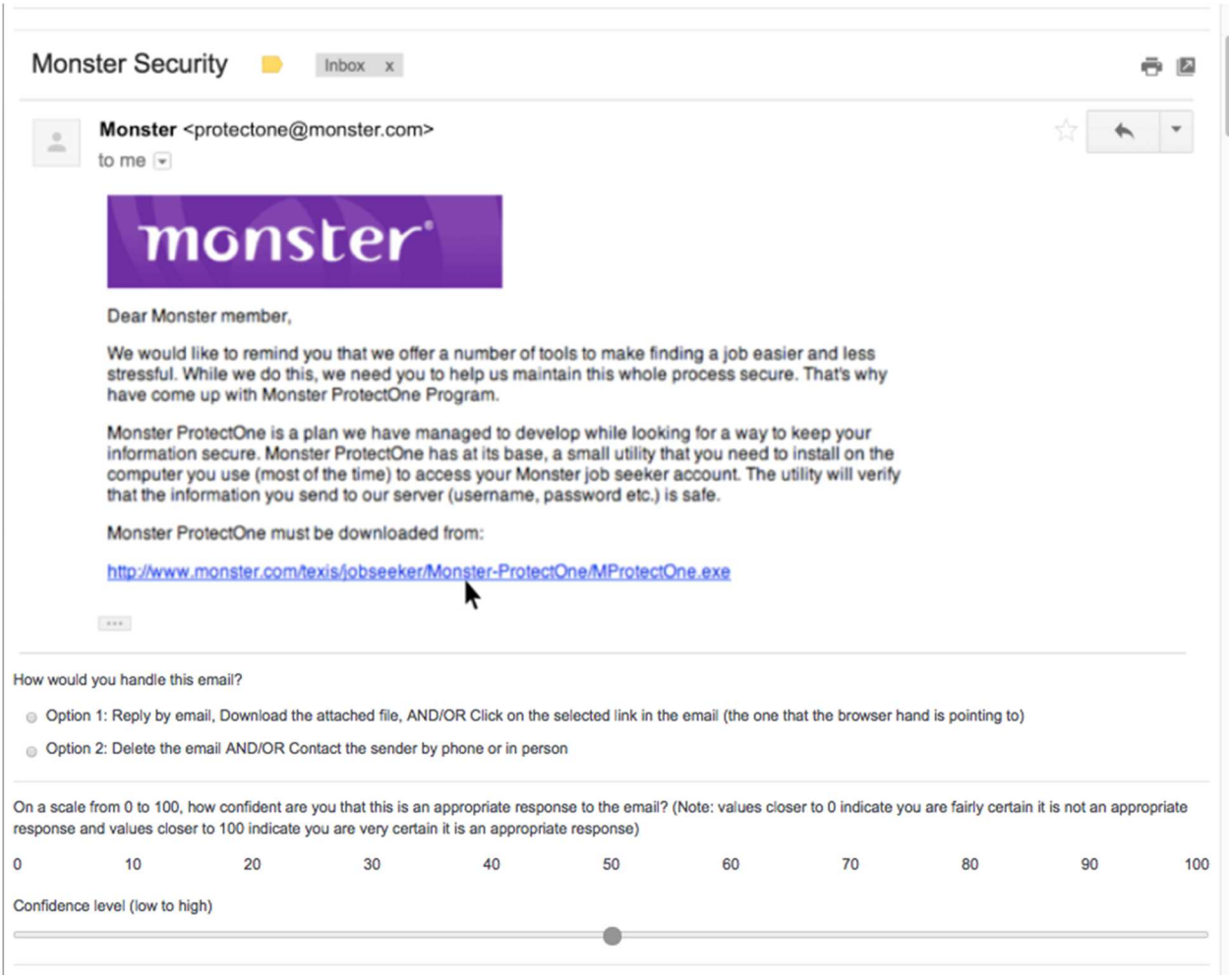
- ☐ Male
- ☐ Female

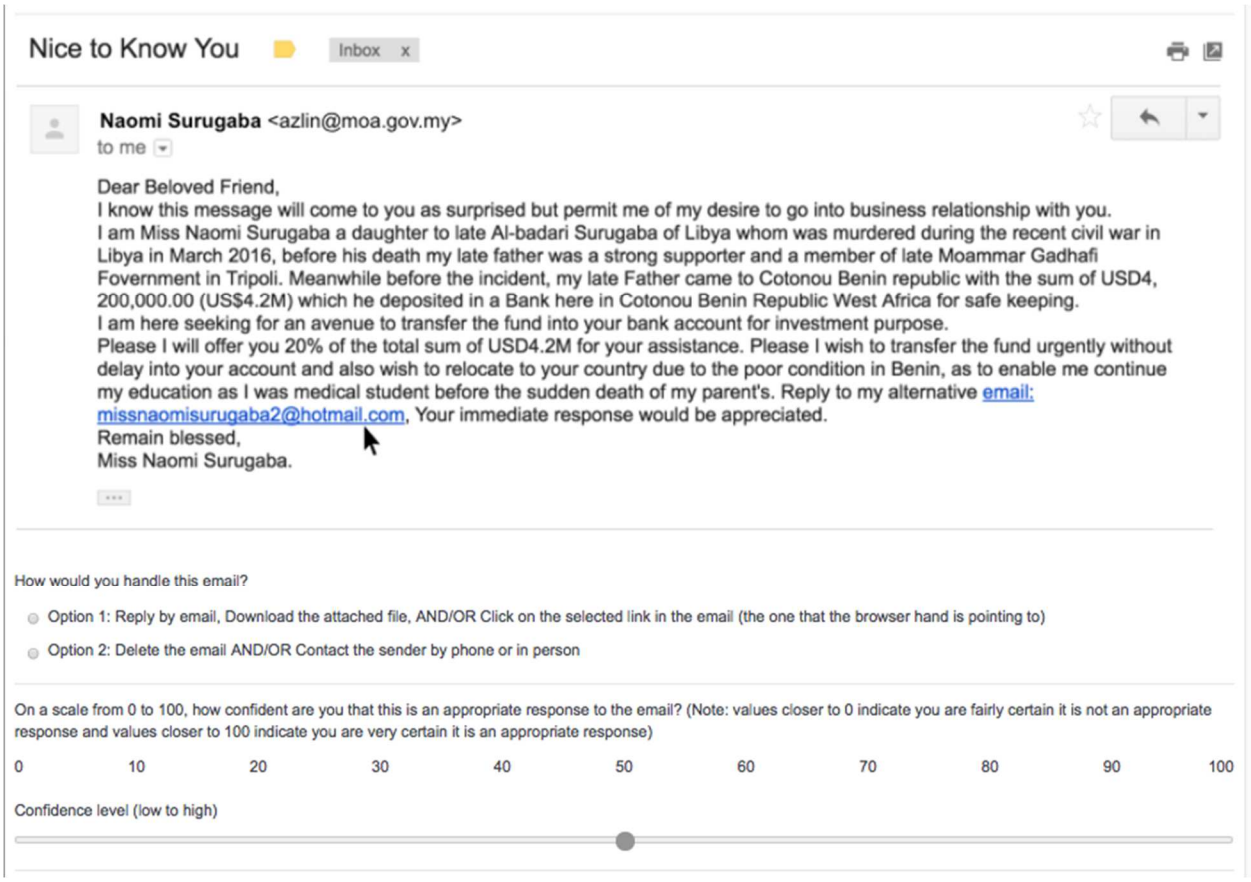
What is your age (in years)? _____

On average, how many hours do you work per week? _____

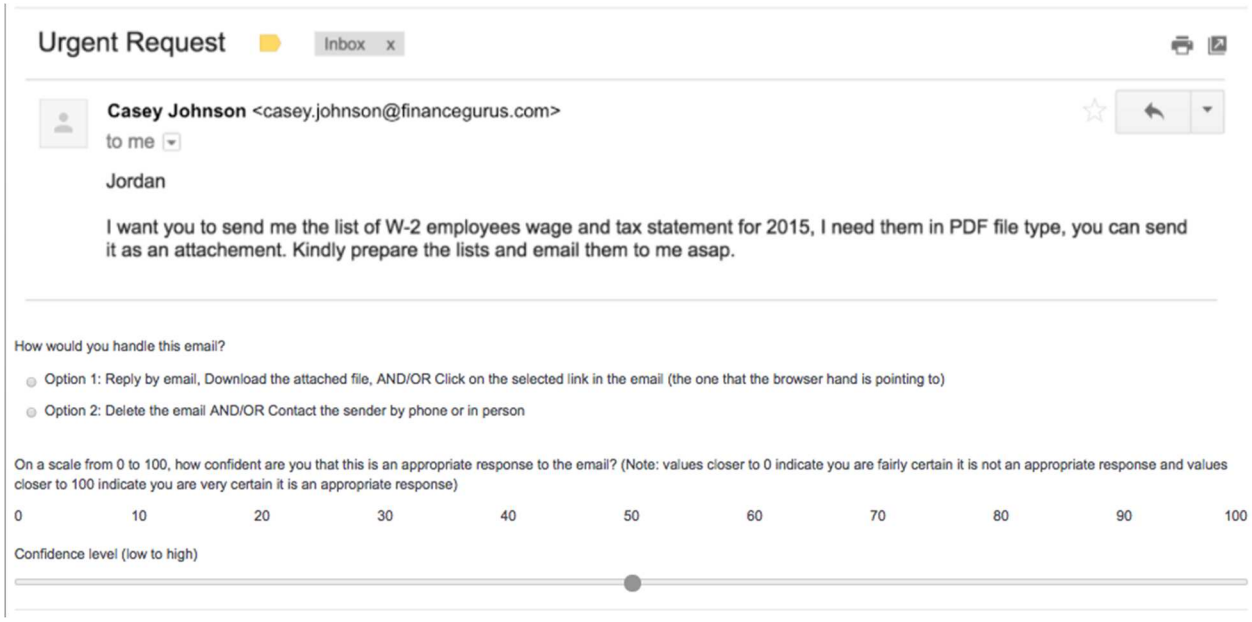
What is your job title? Select*
(*this response option branches through job trees from O*NET classification)

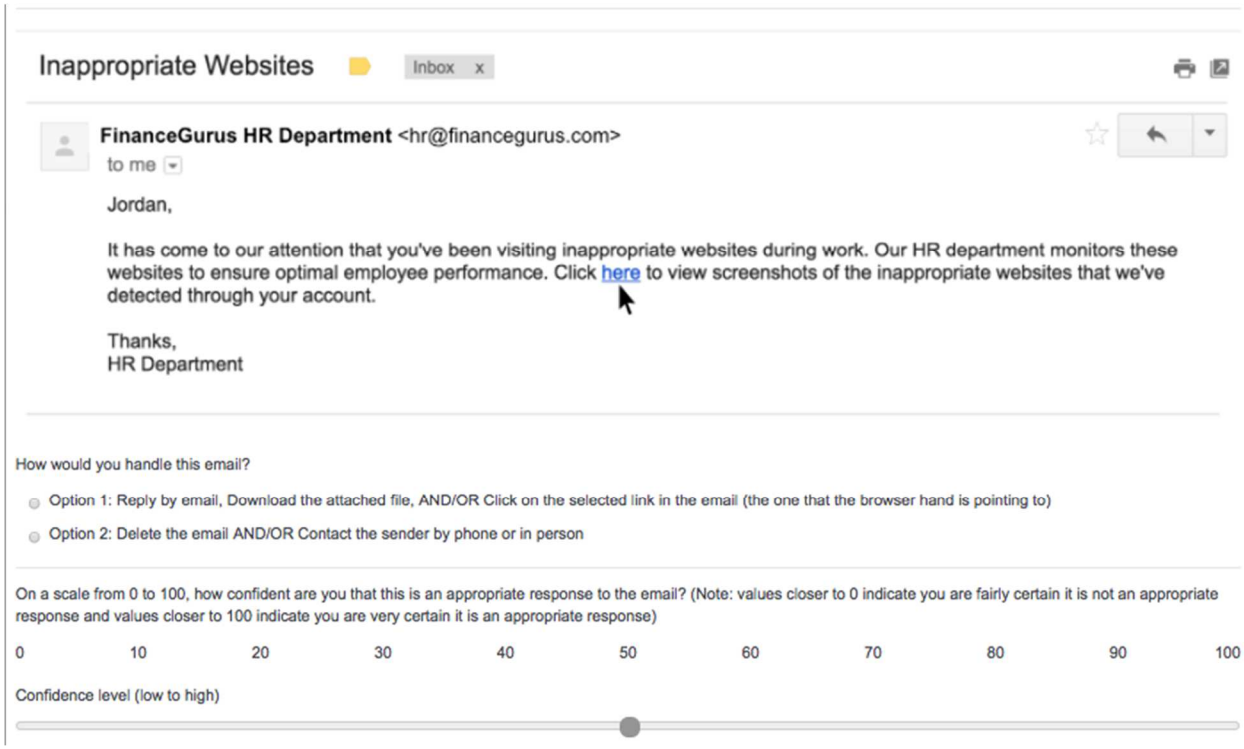
Appendix C: Phishing email examples





Appendix D: Spear-phishing examples





Appendix E: Prompt

Please read the following scenario in order to familiarize yourself with your role for this study.

You are a recruiter, named Jordan Smith, for a small government contracting company called FinanceGurus. Your role involves finding, corresponding with, and evaluating incoming talent, managing the LinkedIn, Taleo, and Monster profiles for the company, and coordinating with the other Human Resources personnel (Chloe) and external hiring services (e.g. Scout, Happie). You are also involved in some administrative work (e.g. company shipping) since it's a small company. Currently, you're sourcing for a candidate for a Senior Accountant position, a SharePoint position and Senior Financial Specialist position. Please read through the following 60 emails and, based on the given information, evaluate the top candidates to suggest to the HR lead, Casey Johnson, for hire.

Appendix F: IRB Approval Letter

USF
UNIVERSITY OF
SOUTH FLORIDA

RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX (813) 974-7091

January 3, 2017

Jaclyn Martin Psychology Tampa, FL 33612

RE: **Exempt Certification** IRB#: Pro00025966 Title: Computer Behavior at Work

Dear Ms. Martin:

On 1/3/2017, the Institutional Review Board (IRB) determined that your research meets criteria for exemption from the federal regulations as outlined by 45CFR46.101(b):

(2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

As the principal investigator for this study, it is your responsibility to ensure that this research is conducted as outlined in your application and consistent with the ethical principles outlined in the Belmont Report and with USF HRPP policies and procedures.

Please note, as per USF HRPP Policy, once the Exempt determination is made, the application is closed in ARC. Any proposed or anticipated changes to the study design that was previously declared exempt from IRB review must be submitted to the IRB as a new study prior to initiation of the change. However, administrative changes, including changes in research personnel, do not warrant an amendment or new application.

Given the determination of exemption, this application is being closed in ARC. This does not limit your ability to conduct your research project.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

Kristen Salomon, Ph.D., Vice Chairperson USF Institutional Review Board