2020

Does One Bad Phish Spoil the Whole Email Load?: Exploring Phishing Susceptibility Task Factors and Potential Interventions

Dawn Sarno
University of Central Florida

DOES ONE BAD PHISH SPOIL THE WHOLE EMAIL LOAD?: EXPLORING PHISHING SUSCEPTIBILITY TASK FACTORS AND POTENTIAL INTERVENTIONS

by

DAWN MARIE SARNO
B.S. Bridgewater State University, 2015
M.A. University of Central Florida, 2018

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Spring Term
2020

Major Professor: Mark B. Neider

ABSTRACT

Phishing emails have become a prevalent cybersecurity threat for the modern email user. Research attempting to understand how users are susceptible to phishing attacks has been limited and hasn't fully explored how task factors influence accurate detection. Even further lacking are the existing training interventions that still have users falling victim to up to 90% of phishing emails following training. The present studies examined how task factors (e.g., email load, phishing prevalence) and a new form of intervention, rather than training, influence email performance. In four experiments, participants classified emails as either legitimate or not legitimate and reported on a variety of other categorizations (e.g., threat level). The first two experiments examined how email load and phishing prevalence influence phishing detection. The third experiment examined the interaction of these two factors to determine whether they have compounding effects. The last experiment investigated how performance can be improved with a novel cheat sheet intervention method. All four experiments utilized individual difference variables to examine how cognitive, behavioral, and personality factors influence detection under various task conditions and how they impact the utilization of training interventions. The results across the first three experiments indicated that both high email load and low phishing prevalence decrease email classification accuracy and sensitivity. However, performance was poor across all conditions, with phishing detection near chance performance and sensitivity values indicating that the task was very challenging. Additionally, participants demonstrated poor metacognition with over confidence, low self-reported difficulty, and low perceived threat for the emails. Experiment 4's results indicated that phishing detection could be improved by 20% with the embedded cheat sheet intervention. Overall, the present studies suggest that email

load and phishing prevalence can decrease fraud detection, but that embedded phishing tips can improve performance.

This dissertation is dedicated to the late Dr. Brendan J. Morse. His mentorship provided me with the skills to continue my academic endeavors and without his support this dissertation would not have been possible.

ACKNOWLEDGMENTS

I am immensely grateful for the support I have received throughout my dissertation process. I would like to thank my advisor, Mark Neider, for his constant encouragement and guidance. He has truly shaped my academic career and I could not be more thankful for the opportunity to work with him. In a similar vein, I would like to thank my dissertation committee for their combined impact on this project. Thank you Dr. James Szalma, Dr. Joseph Schmidt, and Dr. Mindy Shoss, I am so appreciative of your time and thoughtful feedback.

Finally, I would like to thank my family and husband. My older brother JT, has been a constant example of perseverance and hard work. You inspire me to keep going even when I can't see through the darkness. To my husband, who has been on this journey with me since the beginning and witnessed every challenge in this process. Thank you for being my rock during the turbulent experience of graduate school. Lastly, I want to thank my parents. You have supported me when it didn't seem like this would ever be possible. I want to thank you for never giving up hope and always believing I am capable of greatness.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION TO PHISIHNG SUSCEPTIBLITY

Cybersecurity attacks have become a pervasive threat in modern society. As technology rapidly expands, corporations and individuals are only becoming more vulnerable to potential cyberattacks. The U.S. Council of Economic Advisers (2018) estimates that malicious cyber activity cost the U.S. economy somewhere between $57 billion and $109 billion in 2016 alone. This financial cost stems from cyber-attacks affecting both private and public systems in the form of "data and property destruction, business disruption (sometimes for the purpose of collecting ransoms) and theft of proprietary data, intellectual property, and sensitive financial and strategic information" (The Council of Economic Advisers, 2018, p. 1). The latter form of attacks has particular importance for the individual user in the context of phishing emails.

Phishing emails can be defined as "email scam(s) that attempts to defraud people of their personal information" (Drake, Oliver, & Koontz, 2004, p. 1). The computer science domain has attempted to prevent these attacks by removing them from users' inboxes with spam filters. Modern techniques to improve the detection of spam filers involve machine learning to discover the typical characteristics of fraudulent emails. For instance, Fette, Sadeh, and Tomasic (2007) found that at least half of the phishing emails contained a "non-matching" URL and were presented in HTML. Additionally, phishing emails often include company logos and links, attempt to create a plausible premise, and require a quick response (Elkind, 2003; Drake et al., 2004; Jakobsson, 2007). These characteristics demonstrate how phishers can spoof reputable companies and trick email users into interacting with an email. If the email includes a legitimate company's information (e.g., logo, URL address) and a plausible premise then email users may be tricked into believing the email came from a trusted source. Additionally, requiring a quick

response, such as “if you do not respond to this email within 24 hours your account will be deleted”, puts time pressure on the email user. This can often lead to users engaging with fraudulent emails when they may not have under less time pressured circumstances (Drake et al., 2004). Although past attempts to prevent phishing attacks relied on eliminating dangerous emails from users’ inboxes (e.g. implementation of spam filters), as phishing attacks are ever evolving it is impossible to completely insulate users. Thus, recent research has attempted to explore how we can improve email users’ abilities to detect phishing emails to prepare for when spam filters inevitability fail.

Phishing email research has primarily fallen into three methodologies, online surveys that collect data about previous phishing experience (e.g., Grimes, Hough, & Signorella, 2007; Sheng, Holbrook, Kumaraguru, Cranor, & Downs, 2010), sending individuals imitation phishing attacks to evaluate their susceptibility (Ferguson, 2005; Vishwanath, Herath, Chen, Wang & Rao, 2011; Vishwanath, Harrison, & Ng, 2016), and laboratory studies that require participants to classify emails as phishing or not (Canfield, Fischoff, & Davis, 2016; Kumaraguru et al., 2007a; Mayhorn & Nieste, 2012; Sarno, Lewis, Shoss, Bohil, & Neider, 2017). Survey data is often limited by retrospective, self-report methods that make deducing an individual’s real-world behavior with emails difficult. Additionally, although imitation attacks are the most ecologically valid method of testing susceptibility, they are frequently limited in power and control. Due to these limitations, most research has attempted to explore phishing vulnerabilities in the laboratory setting.

Phishing studies vary greatly in methodology even in laboratory settings. One major factor that varies is how researchers ask participants to classify emails. For instance, Canfield et al., (2016) asked participants if the email was phishing, whereas Sarno, Lewis, Bohil, & Neider

(in press) asked participants if the email was spam/safe, and Mayhorn & Nyeste (2012) asked participants if the email was trustworthy. The variability in classifications is important because research has demonstrated that users vary in their classifications based on the word framing (Sarno et al., 2017). Specifically, participants rated the characteristics of emails within three categories, spam, authentic, and dangerous. Ratings differed for each category suggesting they each represent distinct qualities of the content of the email (Sarno et al., 2017). Thus, classifications need to be carefully implemented to ensure that they are representative of phishing emails. As classifications and actions may not always coincide, several studies have also examined what action participants would take with various emails. Participants are often posed with multiple action choices they can choose from, such as reply by email (Canfield et al., 2016; Downs, Holbrook, & Cranor, 2006), delete the email (Canfield et al., Downs et al., 2006; Parsons, McCormac, Pattison, Butavicius, & Jerram, 2013), or follow up by phone (Canfield et al., 2006; Downs et al., 2006). Results demonstrate that classifications are typically consistent with the actions chosen, such that when a participant classifies an email as fraudulent they choose safe actions (e.g., delete/ignore). However, this also means that when participants incorrectly classify emails as safe they choose dangerous actions (e.g., replying or clicking a link). Other factors that have been added in the email classification include confidence and threat levels (Canfield et al., 2016). These factors provide a more holistic representation of a participant's interaction with an email.

Signal detection theory (Mackworth, 1948; Green & Swets 1988) has also been applied to better evaluate phishing susceptibility. Canfield and colleagues (2016) utilized signal detection theory to understand phishing susceptibility in their email classification task. Participants evaluated emails of a fictitious person. Results demonstrated that overall sensitivity to phishing

emails was low and participants treated false alarms as costlier than misses in their classifications, indicating they were conservative in classifying emails as phishing. Interestingly this pattern was reversed for actions suggesting that participants chose to be safer with their actions. These findings indicate that although users typically make action choices that are consistent with their classification, they may be slightly safer with their action selections than classifications. Sarno et al., (in press) found similar results when examining decision profiles for older and younger adults when categorizing emails. Consistent with Canfield et al., (2016), younger adults were more liberal in their ratings of emails, rating more emails as safe. Overall accuracy for the task was low and did not differ between age groups. Lastly, Mayhorn & Nyeste (2012) utilized signal detection measures to evaluate the effectiveness of cyber training methodologies. Immediately after training they found fewer phishing email misses for both of their training conditions. Overall, signal detection measures have allowed researchers to more fully categorize participants' vulnerability to phishing emails and the efficacies of training paradigms (i.e., criterion shifts vs sensitivity improvements).

Recently, a cognitive model has been put forward to better understand phishing vulnerability (Vishwanath et al., 2016). The Suspicion, Cognition and Automaticity Model (or SCAM model) of Phishing Susceptibility accounts for the various factors that are involved when determining the authenticity of an email. The authors chose to utilize suspicion because it has shown to be a successful predictor of trust (McCornack & Parks, 1986) and to be a reliable predictor of performance (Levine & McCornack, 1991). The model also includes a cognitive component which details how cyber risk beliefs and information processing may interact to influence suspicion. Specifically, the authors propose that two forms of cognitive processing are responsible for classifying emails, heuristic processing and systematic processing. Heuristic

processing was described as the utilization of simple decision rules or cognitive heuristic triggers by similar cues in the context. Systematic processing was identified as the careful examination of the quality of arguments within a persuasive context. The former type of processing was proposed to be efficient but error prone, whereas the latter type of processing is supposed to lead to more optimal decisions but be time consuming. Cyber risk beliefs are also included in the cognitive component, not only for their direct impacts on suspicion, but for their modulating effects on information processing. The authors propose cyber risk beliefs as "individuals' perceptions about the risks associated with online behaviors" (Vishwanath et al., 2016, p. 6). They hypothesize that if an individual's cyber risk beliefs indicate that their actions have severe consequences then they will be more likely to engage in a systematic processing of the email. The last major component of the model is the concept of automaticity. Vishwanath and colleagues (2016) propose that deficient self-regulation of email habits results in a decreased ability to detect phishing attacks. The authors contend that this deficient self-regulation causes individuals to not compare their cyber actions with what is appropriate. Additionally, the habitual nature of checking emails makes this domain especially prone to inadequate self-monitoring.

The model was tested against two types of email attacks, link and attachment attacks. Even though the model was only tested on one email per attack type, the results suggested several interesting findings. Specifically, that suspicion regarding the legitimacy of the emails occurred more often when individuals viewed their cyber actions as risky and subsequently engaged in a more systematic processing of the email. Additionally, suspicion was also influenced by the email habits of participants, with more email experience leading to a decreased ability to detect the attack. The authors speculated that email users may fail to detect phishing

emails because they are engaging in habitual behaviors that include automatic and shallow processing of emails. Taken together, these results indicate that engaging in habitual email behavior and complacency often leads to phishing susceptibility. This model provides a potential framework for developing new phishing studies that more accurately examine vulnerability to fraudulent emails.

Regardless of methodology, most research exploring phishing susceptibility has demonstrated that human users are vulnerable to phishing attacks. This vulnerability may be magnified under realistic task settings such as low prevalence rates and high email load. Additionally, the current training literature is limited in its efficacy. The following sections will detail how real-world email task environments influence phishing vulnerability, the current available training paradigms, and how individual differences play a major role in susceptibility. From this research a new phishing email paradigm will be proposed, as well as a new intervention framework aimed at improving email classifications.

Prevalence Rates and Phishing Susceptibility

Most cybersecurity research exploring phishing emails implements a 50/50 split between the phishing emails utilized and the legitimate emails utilized (Canfield et al., 2016; Parsons et al., 2013, Sarno et al., in press). However, the real world rate of phishing emails relative to legitimate emails is estimated to be less than 1% (Canfield et al., 2016). The visual search (e.g., Wolfe, Horowitz, & Kenner, 2005), vigilance (e.g., Baddeley & Colquhoun,1969) and automation (Parasuraman, Hancock, Olofinboba, 1997) literatures suggest that the prevalence of a target directly impacts performance, such that rarer targets are often missed. For example, in the visual search domain, rare targets like tumors can be missed in radiological scans. Similarly,

baggage screeners have more trouble finding weapons in bags when they are infrequently present. Both of these applied tasks have incorporated countermeasures (e.g., breaks, response confirmations, simulated targets) to combat the performance decrements associated with low target prevalence. In the context of phishing emails, this suggests that relative to laboratory settings, email users may be poorer at detecting attacks in realistic settings in which few fraudulent emails occur, and that interventions may be required.

Prevalence in Cybersecurity

Sawyer and Hancock (2019) explored how varying the prevalence of fraudulent emails influenced participants' performance. All participants were presented with 300 emails and were able to either download attachments from the email, reply to the email (and upload their own attachments) or report the email as being potentially dangerous. Importantly, participants either saw phishing emails 1%, 5% or 20% of the time. The results were consistent with the visual search (Wolfe et al., 2005) and vigilance domains (Baddeley & Colquhoun,1969), indicating that when phishing attacks were present 1% of the time they are more likely to succeed. The authors contend that this effect was not found due to fatigue, but rather difficulty discerning attacks because they were infrequent. However, this study was limited in its generalizability since phishing emails were always a request from an unauthorized outside email address ending in a specific suffix. Phishing emails can vary greatly in their characteristics, thus a wider variety of phishing emails is necessary to determine how prevalence affects classifications in the real world. Despite its limitations this study was the first to demonstrate that prevalence is an important task factor in phishing susceptibility and suggests that it needs further examination.

Email Load and Phishing Susceptibility

Highly related to the base rate of phishing emails, is the sheer number of emails a user is evaluating during a given time period. Many cyber security studies provide participants with unlimited time to evaluate emails. Most classification tasks, realistic email probes, and survey studies all allow participants to self-pace in their classifications. However, in the real world this is often not the case. It is likely in the real world that email load is higher, either due to constraints like work demands (e.g., I have a meeting in 5 minutes), or self-inflicted time constraints (e.g., I need to go through these emails quickly so I can watch TV). Additionally, most email users aren't thinking of phishing emails on a daily basis, so when they go through their email they may pace themselves very differently compared to when they know they are in a phishing study. Parsons and colleagues (2013) showed evidence for this hypothesis in their study evaluating the impact of instructions on participants. Participants who knew they were in a phishing study took significantly longer to evaluate emails than participants who did not know they were in a phishing study. This difference in time was also related to riskier actions, where participants who took a shorter time to evaluate the emails (i.e., the control group) were more likely to choose riskier actions than participants who took longer to evaluate to the emails (i.e., the informed group). Thus, when individuals are not expecting phishing emails they evaluate more emails in a shorter period of time and make riskier decisions. Whether self-inflicted or otherwise, email load appears to impact email classifications and the risky actions associated with them.

Time Pressure for Email Classifications

Sarno and colleagues (in press) have also shown that limited time to view emails before classifying them influences decision criteria. Specifically, that without time pressure older adults exhibited conservative response behaviors, rating more emails as spam or not safe. However, once older adults were given a shorter period of time to view emails they became unbiased. These results suggest that decreasing the time to view each email, and thus increasing email load, may directly impact the manner in which individuals evaluate emails. This is inherently important to phishing susceptibility because more users are under time constraints when viewing their emails. Email users can't take all day at work to read their emails, and people often answer emails quickly on their phone. Thus, a better understanding of how email load may impact the manner in which individuals classify and respond to emails is required.

Cyber Events in a Vigilance Task

The vigilance literature can also provide insight into how email load may affect performance with aspects of the task such as event rates. For example, Sawyer and colleagues (2014) examined how the number of cyber events affected performance. Participants took the role of a cyber-defender to monitor strings of IP addresses and communication port numbers on a computer display. Their results indicated that event rate is inversely related to correct threat detection, meaning fast event rates result in decreased threat detection. This result suggests that high email load, where individuals are required to deal with several emails quickly, may result in decreased performance. Intuitively this makes sense, if an individual has less time to evaluate a set of emails, they will be forced to process each email more quickly and superficially. Additionally, Sawyer et al., (2014) found that within the fast event rate condition the probability

of a signal was positively related with correct detection. This finding suggests that in the phishing domain there may be an interesting interaction between the number of emails examined (e.g., email load) and the probability (or prevalence) of a phishing email. Based on Sawyer and colleagues' (2014) results, one might predict that the more emails an individual has the worse they will perform, and that email users will be especially vulnerable to attacks when they are low in prevalence.

Realistic Email Attacks under Varying Email Load

Vishwanath and colleagues (2011) examined how email load may impact phishing performance directly. Participants in their study were targeted with two phishing attacks that were a couple of weeks apart. Although the researchers did not manipulate email load specifically, they collected self-report data from the participants regarding the average amount of email they receive on a given day. Interestingly, email load was not related to how an individual attended to specific cues within an email. However, email load was related to how likely an individual would fall for the phishing attack. Specifically, the more emails participants reported having in their inbox the more likely they were to fall for the phishing attacks. These results suggest that email load is an important aspect of phishing susceptibility, but that further empirical work is required to understand why. It is possible that when email load is manipulated in a controlled setting, clearer findings will be revealed.

Cybersecurity Training and its Limitations

An obvious solution to phishing susceptibility is to train users to be more knowledgeable about the characteristics of phishing emails and the associated risks of interacting with them. Cybersecurity training in general has been shown to decrease, but not eliminate, phishing

vulnerability. Email specific training has been implemented in several studies (Kumaraguru et al., 2007; Mayhorn & Nyeste, 2012; Sawyer et al., 2015). However, many of the studies either have problems in terms of assessing transfer (Sawyer et al., 2015) or poor retention of training (Mayhorn & Nyeste, 2012). Even studies that see robust training benefits compared to controls seem to have difficulty getting participants to stop interacting with phishing emails all together (Kumaraguru et al., 2007b). The existing training methodologies geared towards phishing emails are discussed below, as well as suggestions for potential alternative interventions.

Specific Training and Transfer

Sawyer and colleagues (2015) evaluated how cyber defense training impacted performance in a novel email testbed. Participants were asked to rate 300 emails, with 1% of the emails being fraudulent. Half of the participants received basic cyber-defense training prior to rating the emails. The basic cyber-defense training consisted of a single PowerPoint slide of information (e.g., phishing attempts are fraudulent requests for personal information). Results indicated that even brief training improved performance relative to no training for detecting phishing emails (79% hits, 43% hits, respectively). It is important to note that even though the benefits to performance were large, individuals still missed a large portion of fraudulent attacks (~21%). Indicating that errors still occur with this type of training and any phishing error can result in identity theft. Additionally, the training was extremely specific to non-company email addresses, unsafe attached files, obvious spelling errors and requests for personal information. Any email attacks included this information, thus limiting the generalizability of the findings. Phishing emails can often include embedded links with other indicators such as time pressure, and disproportionate benefits to the sender (Drake et al., 2004). Thus, specific training and

testing of this sort does not extend well to general email performance. If you only train users on a small set of concepts and then test them immediately on those same characteristics, you would always expect to see large improvements. This training methodology also did not explore how participants retained the information provided. It is possible that a delay of only a few hours would result in the information being lost.

Limits in Training Performance

Even more robust cybersecurity training interventions have proven problematic. Kumaraguru et al., (2007b) conducted a study that examined the training efficacies of either embedded training or non-embedded training. The former consisted of immediate descriptive feedback following clicking a link within a fraudulent email. The latter consisted of sending the same information in a secondary email. Their initial results indicated that embedded feedback is more successful in reducing phishing vulnerability immediately following training, benefits users up to one week later, and can even transfer to novel emails. This suggests that the method of training is just as, if not more, important as the information contained within the training. However, even immediately following training, participants were still only at 68% accuracy for detecting phishing emails in the embedded condition, with a slight dip in performance a week later. The non-embedded condition was even worse, with only 14% accuracy for phishing emails immediately following training and 7% accuracy a week later. These results indicate that even immediately following training, performance can still be extremely poor for phishing emails.

Retention of Training

Retention of training is a critical component of any intervention. The previous cybersecurity training methodologies that have been proposed are not successful in terms of retention. For instance, Mayhorn and Nyeste (2012) found benefits of training from both a comic or video immediately following the training. However, two weeks after training, individuals were more likely to fall for a phishing attack than they were before training. Even individuals who have extensive cybersecurity training have been found to be susceptible to phishing attacks quickly after training. The West Point Carronade (Ferguson, 2005) was an investigation that the university conducted after several incidents where cadets were clicking on suspicious attachments and embedded links. The goal of the investigation was to increase awareness and improve cybersecurity performance. A fraudulent email was sent to a random selection of cadets in order to test the susceptibility of the group. Overall, a shocking 80% of cadets clicked on the fraudulent link in the email. Even within four hours of a computer security instruction 90% of the freshmen cadets clicked on the embedded link in the email. Taken together, these two examples emphasize how poor the current retention of training is in the cybersecurity domain. Additionally, given how "successful" training in this domain still exhibits a disturbingly high success rate of phishing attacks, drastically different interventions may be necessary. It is entirely possible that training alone is not sufficient in this domain, and persistent intervention methods need to be taken.

Possible Interventions

Byrne and colleagues (2016) examined self-evaluations of risk and enjoyment while utilizing the internet. Participants were asked to list the types of actions they take on the internet

and why they take those actions, as well as the perceived risk of engaging in those actions. Overall, their results demonstrated that similar to phishing emails, internet users have poor ability to correctly determine risk involved with their online actions. The authors suggest that organizations and individuals can create cheat-sheets that provide an easy and simple view of safe internet behavior. For example, a cheat-sheet could categorize actions by potential risks (i.e., low, moderate, high) or supply suggestions for what to do or what not to do. Within the context of phishing emails, it is possible that users may benefit from an intervention where they are given a cheat-sheet of the typical characteristics of fraudulent emails. This cheat-sheet may be able to guide them in their classification of the email. For example, many phishing emails require a quick response or contain grammar/spelling mistakes (Drake et al., 2004). Thus, a cheat-sheet could include that information stating that if the email contains both grammar/spelling mistakes and a quick response, then it is likely a dangerous email.

One key limitation to learning in the cybersecurity domain is immediate feedback on performance. Schmidt & Bjork (1992) explain why immediate feedback is important to the learning process in terms of encouraging correct behavior and increasing efficient behavior; frequent feedback can also be detrimental to performance. When feedback is too frequent it can become intertwined with the task and when participants do not have the feedback in the test phase they have more problems. While this is a valid concern, Kumaraguru et al., (2007a) demonstrated that immediate and embedded feedback did improve performance relative to the same information presented in an email compared to no feedback at all. Thus, general feedback or embedded information in the email domain may be beneficial. While it is impossible to provide accurate feedback for real phishing emails in the real world, it is possible for email systems to provide periodic feedback or information regarding phishing emails in general. For

instance, while users are engaging with emails they can be presented with pop-ups that include information regarding the dangers of phishing emails or the common characteristics of dangerous emails. This type of method can be implemented in the laboratory and users could continue to utilize it in the real world.

Individual Differences in Phishing Susceptibility

The previous sections have demonstrated that all individuals seem to struggle with correctly classifying and interacting with phishing emails, even with training. However, as with most tasks, there are several individual difference variables that may exacerbate performance decrements in the email domain. For instance, age has been a factor heavily investigated as a potential variable impacting identification of email scams (Grimes et al., 2007; Kircanski et al., 2018; Sarno et al., in press; Sheng et al., 2010). Factors such as age, previous cyber experience, general cyber hygiene, deficient self-regulation, and personality characteristics will be discussed as they relate to phishing vulnerability and the probability of successful interventions.

Younger Adults

Despite previous research suggesting that older adults are a vulnerable age group for phishing attacks (Grimes et al., 2007; Kircanski et al., 2018), several recent studies have indicated that younger adults may actually be more vulnerable to fraudulent emails (Cain, Edwards, & Still, 2018; Sarno et al., in press; Sheng et al., 2010). Sarno and colleagues (in press) examined how phishing vulnerability varies across the lifespan by asking both older and younger adults to classify emails. All participants were required to classify 100 emails (50 phishing, 50 legitimate) either as spam/not spam or safe/or not safe. Despite the fact that overall accuracy did not differ between the two age groups, signal detection measures did highlight important age

differences. Younger adults were found to be more liberal in their ratings, in that they were less likely to classify an email as spam or not safe. Due to this response bias, younger adults missed more phishing emails than older adults. Additionally, although not significant, younger adults had a lower sensitivity for the task than older adults.

The results from Sarno et al., (in press) are particularly interesting when considered in conjunction with several other studies. The West Point Carronade study (Ferguson, 2005) was conducted with young adults, and their presumably youngest group of students (i.e., freshmen) had the worst performance. Additionally, Sheng et al., (2010) discovered that participants ages 18-25 are the most susceptible to phishing attacks. A possible explanation for this may be that younger adults are more comfortable with the typical deception triggers (e.g., spelling mistakes, abnormal structure) in phishing emails from texting. Younger adults may also be more susceptible to fraudulent attacks because younger adults tend to behave less securely online than older adults (Cain et al., 2018). Overall, the previous research indicates that younger adults are an important age group to examine phishing vulnerability due to their decreased ability to detect fraudulent emails and unsafe online behaviors.

Cyber Experience

Individual differences in cyber experience have been heavily examined in the context of cyber-attacks. For example, Silva, Emmanuel, McClain, Matzen, and Forsythe (2015) compared performance between novice and expert cyber incident reporters. Utilizing eye movements, the researchers determined that novice reporters took longer to locate the primary region of interest and were more readily distracted by erroneous text in the display compared to their expert counterparts. These results are indicative of the inefficiency of novices and their decreased

ability to even detect cyber threats. Overall, Silva and colleagues (2015) demonstrated that experience with cyber threats can be predictive of attention and performance in the cyber domain.

Several phishing studies have also examined the relationship between cybersecurity experience and phishing vulnerability. Most research has suggested that more experience tends to lead to more secure online behaviors. For instance, Sheng and colleagues (2015) found that previous experience with cyber educational materials decreased the tendency to enter personal information into phishing websites by 40%. Additionally, Grimes et al., (2007) found that attitudes towards spam email was related to computer expertise, with more negative attitudes associated with higher levels of expertise. These studies suggest that more cyber experience should result in decreased susceptibility to cyber-attacks. However, there has been some research that has suggested that experience may prove detrimental to email classification performance. Specifically, Parsons and colleagues (2013) found that participants who had received formal training in information systems performed worse compared to those who had none. Cain et al., (2018) also found that participants who rate themselves as experts report less secure online behaviors than their more novice counterparts. It is possible that there is a non-linear relationship between experience and cybersecurity performance. Such that those individuals who have nominal training perform the worst because they have a false sense of security, and that true experts exhibit safer and more accurate performance in the cyber domain. Most of the studies exploring cyber experience use one or just a few questions to evaluate previous cyber experience. A more in depth scale investigating previous experience may illuminate the disparate findings in the literature.

Cyber Hygiene

Various aspects of cyber behavior have been linked to phishing susceptibility. Vishwanath and colleagues' SCAM model (2016) indicated that cyber risk beliefs are a predictor of the ability to detect phishing emails. Specifically, that individuals who have an accurate mental representation about the risks of cyber actions are less likely to engage with phishing emails. However, Downs et al., (2006) suggested that general risk awareness may not be connected to an individual's ability to correctly detect phishing emails, rather, that users can only correctly detect phishing emails when they have specific experience with the risks associated in the emails. When users are presented with unfamiliar risks they are still susceptible to phishing attacks, regardless of their awareness of general cyber risks/threats. Overall this study emphasizes the importance of relevant experience as a predictor of performance. It also suggests that specific risky cyber behaviors may play an additional role in phishing detection. Cain and colleagues (2018) examined this idea in the form of cyber hygiene. Cyber hygiene consists of safe online practices, for instance updating your software, using firewalls, anti-virus scans, and not opening emails or attachments from unknown sources. As previously stated, they found that participants who self-identified as experts reported less secure behaviors than their more novice counterparts. Additionally, self-reported experts appeared to have less knowledge about cyber hygiene than other participants. Both findings suggest that cyber hygiene may be distinct from cyber experience, such that experience does not always predict behavior. Additionally, consistent with the training literature, participants who have received cybersecurity training did not exhibit better cyber hygiene, suggesting that training does not improve cyber hygiene at all. Thus, general cyber hygiene may represent a distinct individual difference that describes how users

may interact with phishing emails regardless of previous cyber experience and should be further explored.

Deficient Self-Regulation

Vishwanath and colleagues' SCAM model (2016) suggested that deficient self-regulation is a critical aspect of developing suspicion for fraudulent emails, such that more deficient self-regulators are less likely to develop suspicion. In the SCAM model deficient self-regulation was defined by 8 self-report introspective questions (e.g., I feel my email use has gotten out of control). Other aspects related to deficient self-regulation have been strongly linked to phishing susceptibility in other studies, such as impulsivity. Impulsivity is thought to influence individuals' decisions because those who are impulsive tend to act without reflection and may not attend to predictors of risky behavior (Coutlee et al., 2014). Kumaraguru et al., (2007a) was the first to determine that impulsivity was linked to email classification. Interestingly they didn't find that impulsive individuals were more susceptible to phishing attacks, but rather that less impulsive participants were more likely to engage with emails from companies with which they did not hold an account with. Their explanation was that less impulsive individuals are more reliant on experience and when they don't have mental models of a spoofed company they are more vulnerable to attacks. However, their impulsivity measure was the Cognitive Reflection Test (CRT; Frederick, 2005). This measure utilizes math-based questions that are specifically designed to have an impulsive, but incorrect answer. Correct answers are achieved by thinking more deeply about the problem. The issue with this impulsivity measure is that experience with math may confound the results, such that those who score poorly may also be bad at math rather

than just impulsive. A more sensitive and extensive measure of impulsivity may reveal a relationship with phishing susceptibility.

Interestingly, a different study examining phishing susceptibility and impulsivity found the CRT to be a significant predictor of phishing detection (Parsons et al., 2013). The goal of this study was to determine whether knowing you were in a phishing study impacted your performance. However, they also discovered that participants who were unaware of the study's nature were significantly better at detecting phishing emails when they rated low on impulsivity. No differences as a function of impulsivity were observed for participants who were informed that they were in a study about phishing. These results suggest that impulsivity can play a critical role in phishing detection when users are not aware phishing emails may be present, as in realistic settings. Additionally, Hadlington (2017) found that impulsivity is linked with risky cybersecurity behaviors. Specifically, that the non-planning factor involved with impulsivity results in riskier cyber behaviors. Together, these studies support the idea that deficient self-regulation, and specifically impulsivity, can play a major role in the accurate detection of cyber-attacks.

A different aspect of deficient self-regulation may be inhibitory control. Mayhorn and Nyeste (2012) utilized the Stroop task as a measure of inhibitory control to understand its relationship with phishing susceptibility. Their results demonstrated that inhibitory control is inversely related to phishing susceptibility, such that those with high inhibitory control are less likely to fall for phishing attacks. Thus, the ability to inhibit irrelevant information seems to be a crucial aspect of accurate email classification. Wang, Chen, Vishwanath, and Rao (2012) found supporting evidence for the importance of inhibitory control. Specifically, individuals performed poorly in phishing detection when they attended to visceral triggers over deception triggers.

Phishing emails often contain information that elicits an emotional response (i.e., a visceral trigger), for instance, your bank account may be deleted if you do not respond. When users attend to this information over deception triggers (e.g., spelling mistakes) they are less likely to correctly detect the email as a fraudulent attack. Additionally, Silva et al., (2015) also found that attention to irrelevant information is a critical indicator of cyber performance. Such that novices are often distracted by irrelevant information whereas experts are more likely to attend to the attack relevant information. Overall, it seems that inhibitory control may play a role in phishing susceptibility because individuals who exhibit poor inhibitory control may be unable to ignore information that elicits emotional responses and/or is irrelevant to the task, resulting in an inaccurate (and possibly dangerous) classification of the email.

Personality Factors

Personality is an additional individual difference that may impact vulnerability to fraudulent email attacks. Both conscientiousness and agreeableness have been linked to better cybersecurity performance (McBride et al., 2012, Shropshire, Warkentin, Johnson & Schimidt, 2006; Shropshire, Wakrentin, Sharma, 2015). Conscientiousness typically relates to personality factors such as being careful and thorough (Barrick & Mount, 1991), two personality dimensions that have been previously discussed as important for phishing detection. Indeed, researchers have found that individuals who rate highly as conscientious are less likely to violate security protocols than individuals who rate high on extraversion (McBride et al., 2012). Additionally, agreeableness, although not as directly applicable, can relate to aspects of personalities where individuals are trusting, cooperative, or even compliant (Barrick & Mount, 1991). It is easy to see how being too trusting may impact cyber performance. Email users who are too trusting may

be more likely to fall for fraudulent attacks because of their decreased ability to detect deception. However, agreeableness can also play a role in being compliant or cooperative with cyber protocols that are in place. Importantly, both conscientiousness and agreeableness have been found to relate to compliance with existing security protocols and both the intention and implementation of new security software. Thus, in the context of phishing emails it is likely that individuals who are both conscientious and agreeable may be the most likely to benefit from interventions and training.

Developing Cyber Interventions for Real-World Task Environments

The purpose of the present studies was to examine how phishing vulnerability manifests under real-world constraints. Specifically, how email load and the prevalence of phishing emails impact classification, and also the actions chosen for those emails. Toward that end, all four experiments asked participants to classify each email as legitimate or not legitimate, what action they would take next with each email, and several other classifications (i.e., threat level, difficulty, confidence). Utilizing all of these classifications will determine how classifications/actions function under real world constraints, and if other classifications, such as the perceived threat level of the email or email user's confidence, also differ in more realistic settings. Ultimately, both the classification and actions selected for phishing emails need to be further explored to determine which aspect should be emphasized in future interventions.

Experiment 1 explored how different email loads impact classification and action selection. In a similar vein, Experiment 2 investigated how the prevalence of phishing emails affects classification and action selection. Experiment 3 combined the task factors of email load and phishing prevalence to see, for the first time, how the number of emails and the prevalence

of phishing emails interact to influence classifications and cyber actions. Lastly, Experiment 4 attempted to develop a cheat sheet email intervention that improves performance under high email load and low phishing prevalence.

Hypotheses

Considering the previous cybersecurity research, the following outcomes were expected.

1. Experiment 1: Higher email load should result in the decreased detection of phishing emails and riskier actions taken relative to the lower email load conditions. These effects may be limited by deficient self-regulation and poor cyber hygiene.
2. Experiment 2: Lower phishing email prevalence should cause decreased detection of phishing emails and riskier actions taken compared to higher email phishing prevalence. More cyber experience was expected to attenuate the effect of lower prevalence rates of phishing emails.
3. Experiment 3: High email load and low phishing prevalence was expected to result in the poorest performance, in terms of classification and actions selected. This effect may be limited by deficient self-regulation, poor cyber hygiene and prior cyber experience.
4. Experiment 4: Embedded phishing information and cheat sheets were expected to improve detection for phishing emails under conditions of low phishing email prevalence and high email load. Performance benefits may be limited by individuals who rate low on agreeableness and conscientiousness.

CHAPTER TWO: EXPERIMENT 1

The purpose of Experiment 1 was to determine how email load influences the detection of phishing emails and the actions taken with phishing emails. To my knowledge, no previous research has experimentally manipulated the number of emails participants have to examine within a given timeframe. However, previous research suggests that higher email loads should negatively affect phishing detection, such that having more emails to evaluate results in the decreased detection of fraudulent attacks (Vishwanath et al., 2011).

In order to explore how varying email loads impact email classifications and actions, three groups were utilized. All three groups had unlimited time to view emails and classify emails. Email load was manipulated by changing the number of emails displayed in the inbox. Although all three groups were given 100 emails to classify, some groups were deceived in how many emails they were told they needed to get through. For example, the high email load condition was told they needed to get through 300 emails, and the low email load condition was told they needed to get through 100 emails. Each email was classified for a variety of factors. The main factors of interest included the participants' phishing classification and action selection. Previous research suggests that classifications (i.e., legitimate or not legitimate) are consistent with the actions selected (e.g., delete, ignore) (Canfield et al., 2016; Downs et al., 2006; Parsons et al., 2013). However, it remains unclear whether there is a distinction between classifications and actions when users are put under higher levels of perceived email load. Email users may correctly classify emails under low load but engage in inaccurate and risky actions under pressure.

Other factors of interest that were examined in the context of email classification included confidence ratings, difficulty ratings and threat level assessments. Confidence has been found to relate to email user's sensitivity in detecting phishing emails, but not actions (Canfield et al., 2016). However, again, to my knowledge no one has examined how confidence may be influenced under varying email loads. Specifically, individuals may be more confident when they make risky actions under low email load, but when email load increases confidence may relate to action choice less. Similar relationships were expected for perceived threat level and difficulty ratings.

Deficient self-regulation and cyber hygiene are two individual difference variables that have been previously linked to email performance (Cain et al., 2018; Vishwanath et al., 2016). Deficient self-regulation was measured by two factors, impulsivity and inhibitory control. Impulsivity was measured with the Barratt Impulsiveness Scale Version 11 (BIS-11) (Patton, Stanford, & Barratt, 1995). Each participant's inhibitory control was determined with a Stroop Task (Stroop, 1935). Although impulsivity and inhibitory control have been found to be related to one another (Logan, Schachar, & Tannock, 1997), both measures were included because the impulsivity scale is a more subjective, self-report measure, and inhibitory control is a more objective, direct measure, possibly both representing unique aspects of deficient self-regulation. Individuals who are impulsive and exhibit poor inhibitory control were expected to miss more phishing emails which may limit the effects of email load (Kumaraguru et al., 2007; Parsons et al., 2013; Vishwanath et al., 2016). Additionally, I expected that these individuals would also select riskier actions with phishing emails than individuals who are less impulsive and demonstrate more inhibitory control. Lastly, I utilized 20 questions about cyber practices (e.g., do you secure your browser, do you perform weekly anti-virus scans) from Cain et al., (2018) as

a measure of cyber hygiene. Individuals who demonstrate better cyber hygiene were expected to be more resilient to phishing attacks and exhibit a decrease in the influence of email load.

Method

Participants

Seventy-five undergraduate students (M_{age} = 19.08, 45 males, 30 females) from the University of Central Florida were recruited for course credit. All participants had normal or corrected-to-normal vision (20/32 or better corrected vision on a Snellen eye chart) and color vision (Ishihara's test for color blindness; 13 plates).

An ANCOVA power analysis was conducted in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) in order to determine how many participants would be required to find an effect of email load on performance, controlling for both deficient self-regulation and cyber hygiene. Sawyer et al., (2014) found an effect size of η_p^2 = .47 for event rate in their cyber vigilance task. However, given that the present task is an email classification rather than an IP monitoring task I utilized a smaller and more conservative effect size of η_p^2 = .25 for the analyses to ensure sufficient power. Additionally, deficient self-regulation and cyber hygiene were included as covariates since they have both been found to influence email classification and may account for more variance than the email load manipulation. Thus, I calculated an ANCOVA power analysis with the following parameters, a Cohen's *f* of .58, power of 0.95, an alpha probability of 0.01, 3 groups and 2 covariates. Based off this analysis, 51 participants (17 in each group) should be satisfactory to detect significant differences in email classification.

Apparatus and Stimuli

The experiment was programmed and run in SR Research Ltd's Experiment Builder. Stimuli were real emails, obtained from either the researcher's inboxes/junk folders or web searches and have been utilized in previous studies (Patel, Sarno, Lewis, Neider, & Bohil, in press; Sarno et al., in press; Williams et al., 2019). Participants had unlimited time to view the 100 emails. However, the number of the emails displayed in the inbox depended on the condition (see Figure 1). In the high email load condition participants were told that they needed to get through 300 emails. In the moderate email load condition participants were told they needed to evaluate 200 emails. Lastly, in the low email load condition participants were told they needed to assess 100 emails.

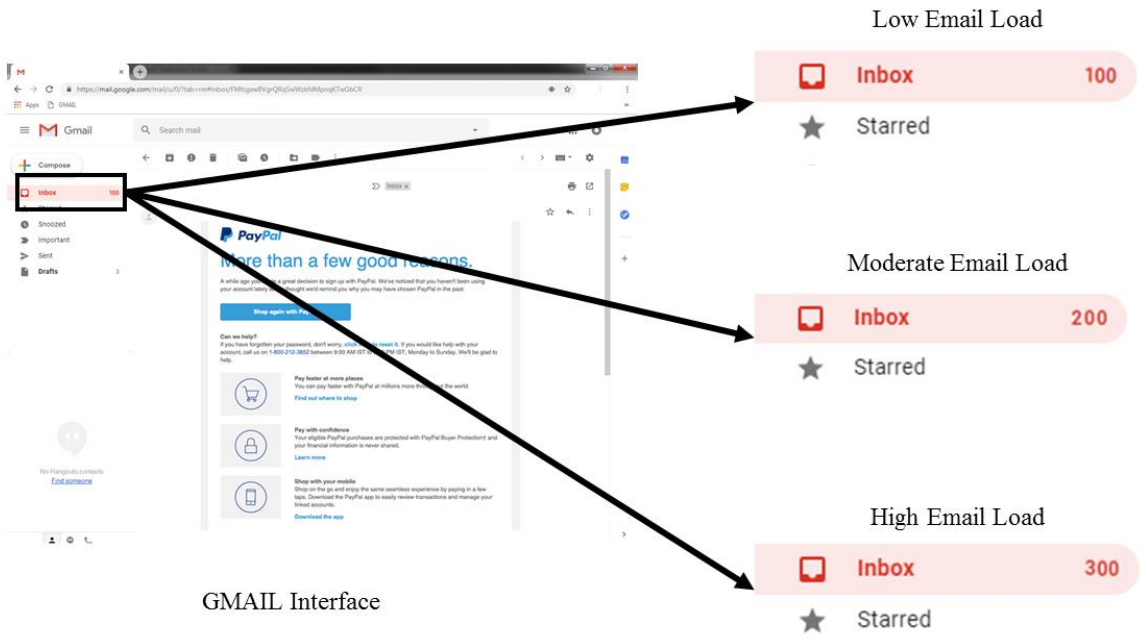


Figure 1.Gmail Interface & Load Conditions

GMAIL Interface with the three email load conditions. Each conditions starts off with a different number of emails displayed in the inbox.

The emails that were utilized were diverse in nature, including content such as banking, media (e.g., Netflix), and shipping (see Figure 2). In order to limit the prevalence effects and have more power, 50% of the emails were real phishing attacks and 50% of the emails were real legitimate emails. The emails were presented within a GMAIL interface that counts down the number of emails in the inbox (see Figure 1). The experiment was presented on a 19" Dell Professional P190S Monitor at a resolution of 1280 X 1040 pixels with participants seated approximately 20 inches away, making the visual angle of the display roughly 36° x 29°. Participants made classifications regarding each email utilizing the mouse and keyboard.

A. B.

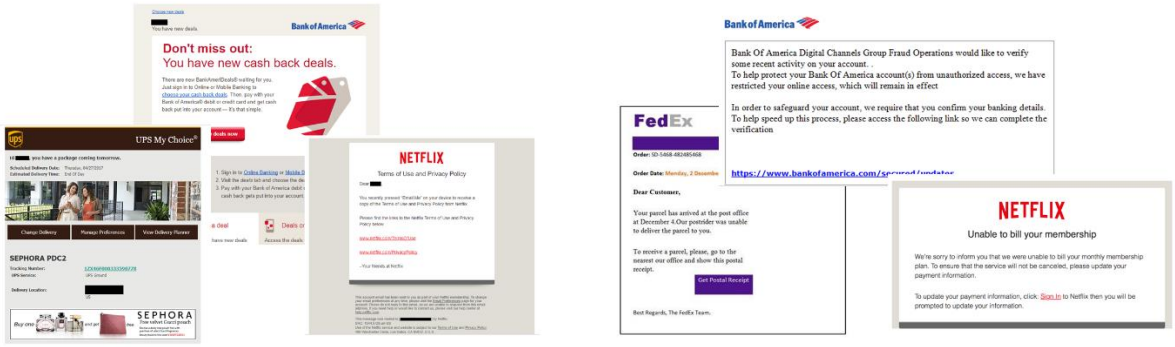


Figure 2. Example emails

Examples of banking, media (e.g., Netflix), and shipping emails. A) Legitimate emails. B) Phishing Emails

Individual Difference Measures

Deficient Self-Regulation

Deficient self-regulation was assessed utilizing two different measures, impulsivity and inhibitory control. Impulsivity was assessed utilizing the Barratt Impulsiveness Scale Version 11 (BIS-11) (Patton et al., 1995). The BIS-11 is a 30-item scale that measures impulsiveness. Items are rated on a 4-point Likert-type scale with anchors ranging from rarely/never to almost always/always. Example items include “I do things without thinking” and “I buy things on impulse.” After reverse scoring, higher scores indicate more impulsive tendencies. Inhibitory control was measured utilizing a Stroop task (Stroop, 1935). The Stroop task consisted of 240 trials where participants were asked to indicate the color of ink a word is written in. The Stroop task is a measure of inhibitory control because in order to respond correctly participants must respond to the color the world is written in and inhibit their response to the word’s meaning. Scores were calculated by taking the difference in response times between congruent trials (e.g.,

the word “yellow” written in yellow ink) and incongruent trials (e.g., the world “yellow” written in red ink).

Cyber Hygiene

Cyber hygiene was measured utilizing the 20 yes/no cyber practice questions from Cain et al., (2018). Example items include “do you secure your browser?” and “do you perform weekly anti-virus scans?” Responding yes to more questions is indicative of better cyber hygiene.

Design

Design of Conditions

The number of emails perceived to be in the inbox depended upon on which condition the participant was assigned, either 300 emails (high load), 200 emails (moderate load), or 100 emails (low load). Half of trials contained a phishing email and half contained a legitimate email. Thus, the overall design of the experiment was a 3 (email load: high vs moderate vs low) x 2 (email type: phishing vs legitimate) mixed factorial design with the first factor between-subjects, and the second factor within-subjects.

Design of Trials

Unique emails were used for each trial. For each of these emails participants were asked to make a variety of classifications (see Figure 3). After viewing an email, participants were first asked to classify whether the email was legitimate or not legitimate via button press. Participants were then asked to rate, on a sliding 5-point Likert-type scale, the threat level of the email,

ranging from not threatening to threatening. Following threat level, participants chose which action they would be mostly like to take next with the email (i.e., click a link or attachment, reply, check sender's address, delete, or report as suspicious). Next, participants completed another sliding 5-point Likert-type scale indicating how difficult their decision was, ranging from not difficult to difficult. Lastly, participants were asked to rate their confidence, also on a sliding 5-point Likert-type scale, from not confident to confident.

Procedure

Upon providing informed consent, participants were prescreened for near, far, and color vision. After being screened for normal or corrected-to-normal vision, participants completed the demographics questions. The demographics questionnaire included questions regarding basic information (e.g., gender, age, education level), questions about their cyber hygiene and the BIS-11 (Patton et al., 1995). After completing the demographics, participants continued to the experimental station in the back of the room for the remainder of the study.

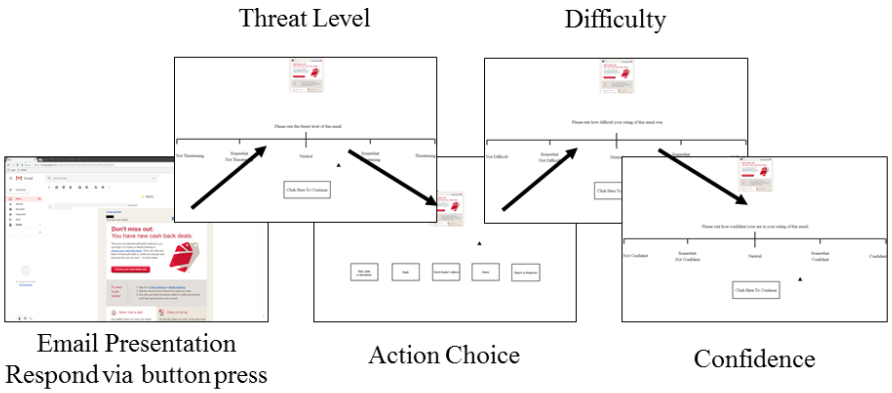


Figure 3. Example Trial Sequence

Each trial began with the email presentation. Then participants classified the email, rated the threat level, chose which action they would take next, rated how difficult their decision was, and indicated how confident they were in their decision.

Prior to completing the experiment participants completed a 240 trial Stroop task programmed with E-Prime. Participants were asked to indicate, via button press, the color of the ink a word was written in. After the Stroop task, participants received the instructions for the experiment. Participants were randomly assigned to one of the three email load conditions (high, moderate or low). Each trial began by presenting an email (see Figure 3). Participants were then asked to indicate via button press if the email was legitimate or not. Participants then classified what threat level the email posed, what action they would take next (i.e., click a link/open attachment, reply, check sender, delete, report as suspicious), how difficult their classification was, and finally how confident they were in their classification (see Figure 3). After completing all of the trials, participants were debriefed regarding the true nature of the study.

Results and Discussion

Data analysis was based primarily on accuracy and response times for the email classifications. Additional analyses were conducted on aspects of the task such as the threat level of the email, what action would be taken next, how difficult was it to assess, and how confident participants were in their responses.

Email Classifications

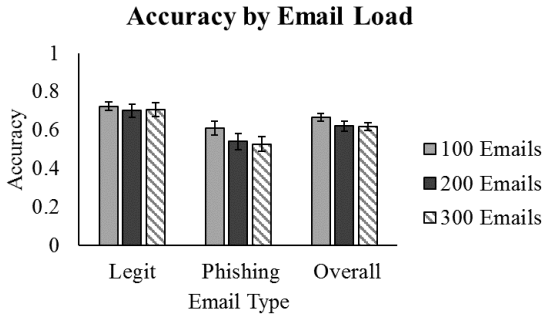
The primary analysis of interest explored how participants accurately classified emails as either legitimate or not legitimate. Neither deficient self-regulation nor cyber hygiene were correlated with any of the dependent measures and therefore were excluded as covariates in all analyses. Accuracy and response times were each submitted to separate two-factor mixed ANOVA with an alpha level of .05, with email load (high, moderate, low) and email type (legitimate, phishing) as the independent variables. Response times were calculated on both correct and incorrect trials. If email load negatively effects the ability to correctly detect phishing emails, then participants who are in the high load condition should be significantly less accurate and faster in their classifications than participants in the moderate and low load conditions.

Email Classification Accuracy

There was a main effect of email type, $F(1,72) = 22.35$, $p < .001$, $\eta_p^2 = .24$, with participants being more accurate in their classifications of legitimate emails (70.88% correct) than phishing emails (55.81% correct) (see Figure 4A). Note that overall the participants were nearly at chance performance for phishing emails. There was not a significant interaction of email type and email load, $F(2,72) = 0.36$, $p = .698$, $\eta_p^2 = .01$, or a main effect of email load,

$F(2,72) = 1.53$, $p = .224$, $\eta_p^2 = .04$, suggesting that email load did not influence the accurate detection of either phishing or legitimate emails (see Figure 4A). Although cyber hygiene was not related to email classifications for legitimate emails, there was a relationship between cyber hygiene and classifications for phishing emails. Specifically, the more phishing emails participants detected the more likely they were to have reported more "hygienic" (i.e., safer) cyber behaviors, r(75) = .26, p = .026. Even though this relationship is relatively weak, it does suggest that general safe cyber behaviors are linked to the ability to detect phishing emails.

A.



B.

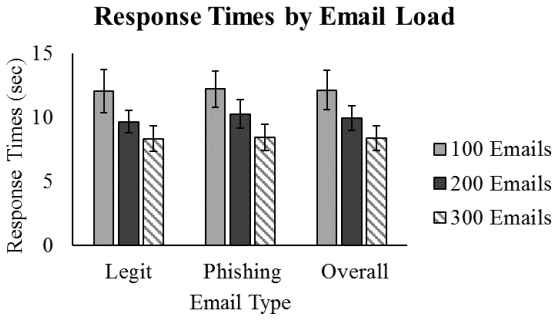


Figure 4. Experiment 1 Email Classification Accuracy (A) and Email Classification Response Times (B) by Email Load and Email Type

Error bars indicate the standard error of the mean.

Email Classification Response Times

There were no main effects of email type, $F(1,72) = 0.89$, $p = .349$, $\eta_p^2 = .01$, or email load, $F(2,72) = 2.47$, $p = .092$, $\eta_p^2 = .06$, nor any significant interaction of the two on response times, $F(2,72) = 0.23$, $p = .792$, $\eta_p^2 = .01$ (see Figure 4B)**.** These results suggest that the time to classify emails does not depend on whether the email is a phishing or legitimate email or how many emails need to be evaluated. However, it is worth noting that there was a positive

relationship between the time it took to classify legitimate emails and phishing detection, r(75) = .31, p = .006, suggesting that there may be a link between classification time in general with phishing detection.

Sensitivity and Response Criterions

Exploring classification accuracy alone may not fully explain performance differences between different email loads. Signal detection measures have been utilized in previous cybersecurity studies (Canfield et al., 2016; Sarno et al., in press) to investigate whether performance differences are due to changes in sensitivity (d') to phishing emails or response criterion shifts (c). Response criterion (c) was chosen over response bias (β) because of the conservative and liberal bounds being more balanced. In response criterion (c) lenient responders have scores that are <0, conservative responders have scores that are >0 than zero, and unbiased responders have scores of 0 (Stanislaw & Todorov, 1999). Response bias (β) scores are limited to 0-1 for lenient responders and can be any number above 1 for conservative responders (Green & Swets, 1988). Both sensitivity and response criterions were calculated for each email load group and subjected separately to separate one-way between subjects' ANOVAs with an alpha level of .05, with email load (high, moderate, low) as the independent variable. Increasing the email load was not expected to change sensitivity but was expected to change the response criterion of the participants. Specifically, participants were expected to be more liberal in their classifications under conditions of higher email load, classifying more emails as legitimate.

There were no main effects of email load for sensitivity (see Figure 5A), $F(2,72) = 1.10$, p =.340, η_p^2 = .03, or for response criterion (C) (see Figure 5B), $F(2,72) = 0.47$, p =.628, η_p^2 =

.01. However, all participants demonstrated very low sensitivities (below 1). Response criterions for each email load condition were each submitted to one-sample t-tests to determine if they were different from zero. Each conditions' average response criterion did significantly differ from zero, (p's > .035) suggesting that all participants were liberal in the responses (i.e., rated more emails as legitimate). Taken together these results suggest that email load does not influence the response profiles of email users, but that all users are very vulnerable to phishing emails.

A. B.

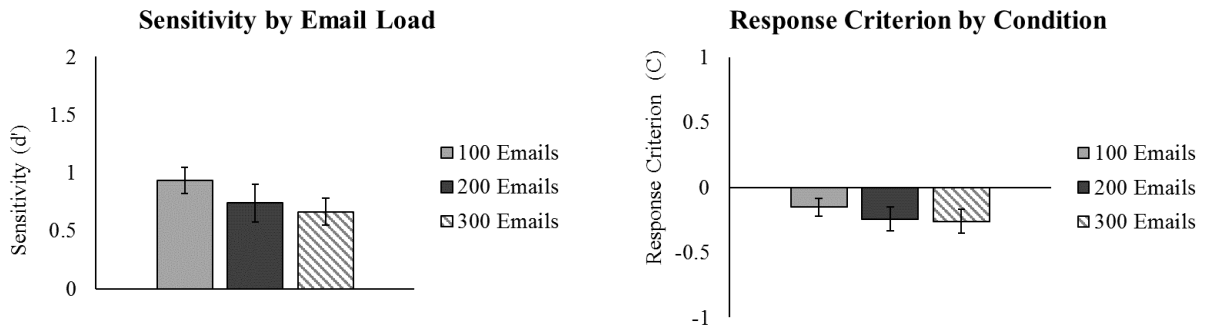


Figure 5. Experiment 1 Signal Detection Measures, Sensitivity (A) and Response Criterion (B) by Email Load

Error bars represent the standard error of the mean.

Actions Chosen

The next actions chosen for each email were also analyzed. Any of the five actions provided may have been acceptable for legitimate emails. In the real world there may be consequences for ignoring legitimate emails. For example, if someone ignores an email from

their credit card company about potential fraudulent activity, the scammer is likely to continue making fraudulent purchases. However, since participants in this study are evaluating emails meant for other individuals it is impossible to know what the correct actions are for legitimate emails. On the other hand, if users reply or click a link in a phishing email they are always putting themselves at risk in the real world, whether it was meant for them or not. Thus, only phishing emails were considered in these analyses. Actions chosen were considered correct for phishing emails if participants chose to check the sender, delete, or report it as suspicious. Incorrect actions for phishing emails included clicking a link/opening an attachment or replying. Action choice accuracy was submitted to a one-way between subjects ANOVA with an alpha level of .05 with email load (high, moderate, low) as the independent variable. Riskier actions (i.e., clicking a link/opening an attachment, replying) were expected to be more prominent in the high load condition if email load negatively impacts email performance.

There was not a main effect of email load on action accuracy (see Figure 6), $F(2,72) = 0.20$, $p = .823$, $\eta_p^2 = .01$, indicating that email load does not meaningfully impact the actions selected for each email.

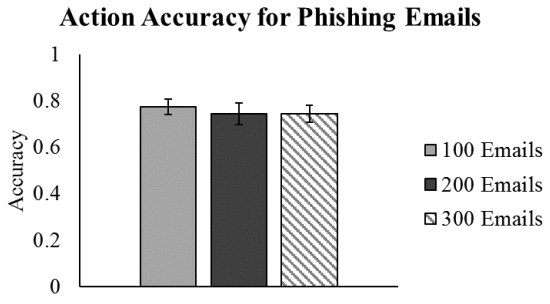


Figure 6. Experiment 1 Action Accuracy for Phishing Emails

Error bars represent the standard error of the mean.

Threat Level, Confidence, and Difficulty

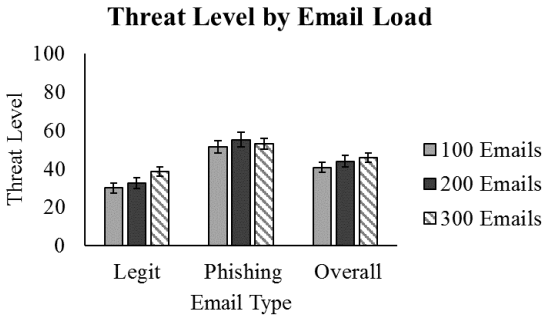
Threat level, confidence, and difficulty may represent different dimensions of the email classification task. Additionally, these relationships may not be consistent across varying email loads. Therefore, three separate two mixed factor ANOVAs, with alpha levels of .05, and email load (high, moderate, low) and email type (legitimate, phishing) as the independent variables were conducted. All three measures were calculated including both correct and incorrect trials. It was expected that higher email loads would result in decreased perceived threat level and confidence, and increased task difficulty due to self-imposed time pressure. However, it is possible that aspects of the overall classification, like confidence, may not vary across email loads, because individuals are unaware that their classifications have become less accurate.

Threat Level

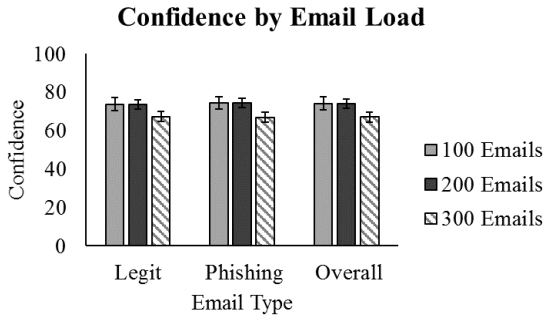
There was a main effect of email type on the perceived threat, $F(1,72) = 155.43$, $p < .001$, $\eta_p^2 = .68$, such that phishing emails were rated as higher threats (53.21) than legitimate emails (33.72) (see Figure 7A). Even though phishing emails were perceived as more threatening they

still were rated rather low on threat level, ~53, out of 100, suggesting that all participants had miscalibrated perceptions of threat. There was not a main effect of email load, $F(2,72) = 0.89$, $p = .417$, $\eta_p^2 = .02$, nor was there an interaction between email type and email load, $F(2,72) = 2.55$, $p = .085$, $\eta_p^2 = .07$, on threat level (see Figure 7A). Overall, these results indicate that email users rate phishing emails as mildly threatening regardless of the number of emails in their inbox.

A.



B.



C.

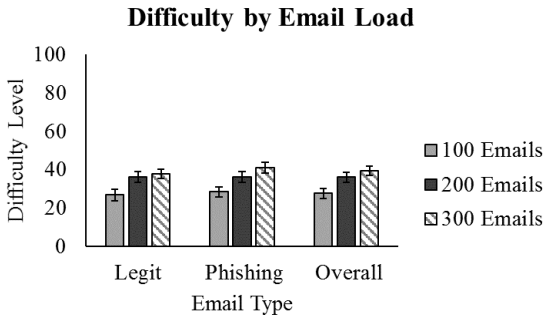


Figure 7. Experiment 1 Threat Level (A), Confidence (B) and Difficulty (C) ratings by Email Load

Error bars represent the standard error of the mean.

Confidence

There were no main effects of email type, $F(1,72) = 0.28$, $p = .596$, $\eta_p^2 = .01$, or email load, $F(2,72) = 2.18$, $p = .120$, $\eta_p^2 = .06$, nor any interaction of the two, $F(2,72) = 0.28$, $p = .760$, $\eta_p^2 = .01$, on confidence in the task, suggesting that confidence is not influenced by the number of emails or whether the email is phishing or legitimate in nature (see Figure 7B). It is worth noting that confidence was fairly high given the relatively poor accuracies across the groups.

Difficulty

There was not a main effect of email type on difficulty, $F(1,72) = 3.50$, $p = .065$, $\eta_p^2 = .05$, suggesting that participants viewed both phishing and legitimate emails as equally difficult (see Figure 7C). There was also not a significant interaction of email load and email type on difficulty, $F(2,72) = 1.22$, $p = .302$, $\eta_p^2 = .03$. However, there was a main effect of email load on difficulty, $F(2,72) = 5.33$, $p = .007$, $\eta_p^2 = .13$, indicating that the number of emails in the participant's inbox influenced how difficult the task was (see Figure 7C). Specifically, pairwise comparisons revealed that when participants were given 100 emails their task was perceived as easier (27.56) than when they were told they had to evaluate 200 emails (36.01; $p = .026$) or 300 emails (39.32; $p = .002$). However, there was no difference between the 200 and 300 conditions ($p = .375$). Taken together these results suggest that the number of emails to examine does influence how difficult the task is perceived rather than the type of email (e.g., phishing or legitimate).

CHAPTER THREE: EXPERIMENT 2

Experiment 1 investigated how perceived email load affects phishing email detection. However, email load is only one task factor involved in email classifications. Experiment 2 explored how the prevalence rate of phishing emails impacts performance. Previous research exploring prevalence rates of phishing emails suggests that when the probability of a phishing email is low, users have poorer phishing detection (Sawyer & Hancock, 2018). However, while Sawyer and Hancock (2018) explored phishing prevalence, they did not specifically investigate the connection between classification and action. Additionally, Sawyer & Hancock (2018) did not utilize an email database with a diverse set of emails that generalizes to most email users (i.e., they utilized only clerical emails). The current study aimed to see if manipulating the number of phishing emails present affects classification, or the next action chosen. Previous research suggests that lower prevalence rates of phishing emails will decrease detection performance, however it remains unclear precisely how this will impact both classifications and actions in a diverse email set.

Method

Participants

Fifty-four undergraduates (M_{age} = 18.65, 19 males, 35 females) from the University of Central Florida participated for course credit. All participants had normal or corrected-to-normal vision and were prescreened for near and far vision (20/32 or better corrected vision on a Snellen eye chart) and color vision (Ishihara's test for color blindness; 13 plates).

In order to determine how many participants were necessary to find an effect of prevalence a new power analysis was conducted in G*Power (Faul et al., 2007). Sawyer & Hancock (2018) found an effect size of η_p^2 = .24, for response accuracy in their three-level prevalence analysis. Additionally, since cyber experience may play a vital role in how prevalence affects accurate detection of phishing emails, I conducted a power analysis considering cyber experiences as a covariate. Thus, I calculated an ANCOVA power analysis using a Cohen's *f* of .56, power of 0.95, an alpha probability of 0.01, 3 groups and 1 covariate. Based off this analysis, 54 participants (18 in each group) should be satisfactory to find signification differences between the three prevalence rates.

Apparatus and Stimuli

The apparatus and stimuli were the same as Experiment 1 with the following exceptions. All participants evaluated 100 emails and were given an accurate email counter. The number of phishing emails depended on condition. The low prevalence (5%) condition contained 5 phishing emails, the moderate prevalence (25%) condition contained 25 phishing emails, and the high prevalence (50%) condition contained 50 phishing emails. Importantly, the same phishing emails in the 5 phishing prevalence condition were utilized in both the 25 and 50 conditions in order to make direct comparisons in performance.

Individual Difference Measure

Cyber Experience

Cyber experience was assessed utilizing 20 self-report questions about an individual's previous experience with cyber threats. These questions were developed for an unpublished

study by Sarno, McPherson, and Neider. Example items include “have you had any previous training about cybersecurity?” and “have you ever had a virus due to engaging with a spam email.” After reverse scoring, higher scores indicate more cyber experience.

Design

Design of Conditions

There were a total of 100 trials, with either 5%, 25%, or 50% of trials containing phishing emails and the remaining trials containing legitimate emails. The overall design for the experiment was a 3 (prevalence: 5% vs 25% vs 50%) x 2 (email type: legitimate vs phishing) mixed factorial design with a between-subjects of prevalence, and a within-subjects factor of email type.

Design of Trials

The design of trials was identical to that of design of trials in Experiment 1.

Procedure

The procedure was the same as Experiment 1 with the following exceptions. Instead of being asked questions about their cyber hygiene participants were asked questions about their cyber experience. Additionally, participants were not given the BIS-11 or the Stroop Task. Lastly, instead of varying the email load, participants were randomly assigned to one of the three prevalence conditions (5%, 25%, 50%).

Results and Discussion

As with Experiment 1, data analysis was based primarily on accuracy and response times for the email classifications. Other analyses were performed on additional aspects of the task, such as the threat level of the email, what action would be taken next, how difficult was it to assess, and how confident participants were in their responses.

Email Classifications

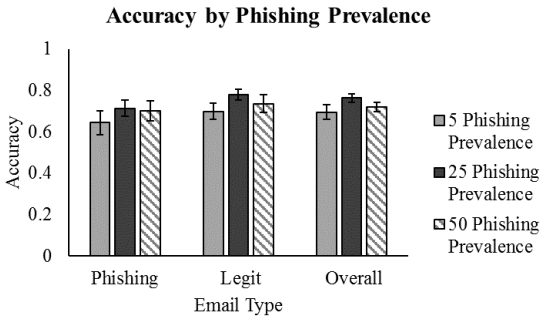
Similar to Experiment 1, the main analysis in Experiment 2 investigated how participants classified emails as legitimate or not legitimate. Cyber experience was not related to any of the dependent measures and therefore was not included as a covariate in any analysis. Accuracy and response times were each subjected to a two-factor mixed ANOVA with an alpha level of .05 with prevalence (high, moderate, low) and email type (legitimate, phishing) as the independent variables. Like Experiment 1, response times were calculated across both correct and incorrect trials. If prevalence influences an email user's ability to correctly identify phishing emails, then classification accuracy for phishing emails should be lowest and fastest when phishing emails are the least prevalent (i.e., the 5% condition).

Email Classification Accuracy

There was not a main effect of email type on accuracy, $F(1,51) = 1.40$, $p = .242$, $\eta_p^2 = .03$, suggesting that participants classified phishing and legitimate emails equally well (see Figure 8A). There was also not an interaction between email type and prevalence, $F(2,51) = 0.04$, $p = .961$, $\eta_p^2 < .01$. There was a marginal main effect of prevalence, $F(2,51) = 3.02$, $p = .058$, $\eta_p^2 = .11$, indicating that prevalence may influence email classification accuracy. Pair wise

comparisons revealed that this difference was driven by comparing the 5 phishing email prevalence condition (67% accuracy) with the 25 phishing email prevalence condition (75% accuracy) (p = .019) (see Figure 8A). There were no other differences amongst the groups (p's >.134). These accuracy measures may be misrepresentative of the true differences in classification accuracy for the prevalence conditions since each group evaluated fundamentally different email sets (i.e., different number of phishing emails). Thus, additional analyses were conducted on the same five emails that each prevalence condition received. This analysis suggested that there were no significant differences amongst the prevalence groups when directly comparing the same 5 phishing emails (see Figure 9A), $F(2,51) = 1.14$, $p = .327$, $\eta_p^2 = .04$. Together these results suggest that lower phishing prevalence may result in poorer overall email classifications.

A.



B.

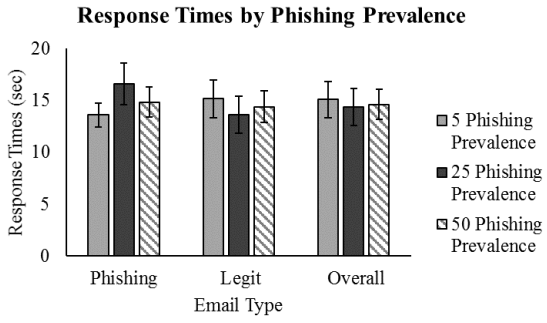


Figure 8. Experiment 2 Email Classification Accuracy (A) and Email Classification Response Times (B) by Phishing Prevalence and Email Type

Error bars represent the standard error of the mean.

Email Classification Response Times

There was no main effect of email type on classification times, $F(1,51) = 0.73$, $p = .396$, $\eta_p^2 = .01$, indicating that the time to classify phishing and legitimate emails was the same (see Figure 8B). There was also not a main effect of prevalence on classification times, $F(2,51) = 0.06$, $p = .940$, $\eta_p^2 < .01$, suggesting that the prevalence of phishing emails does not influence classification times. There was a significant interaction between email type and phishing prevalence, $F(2,51) = 3.21$, $p = .049$, $\eta_p^2 = .11$. Separate one-way repeated measures ANOVAs on phishing and legitimate classification times for each group revealed that this interaction was driven by the 25 prevalence condition. Specifically, in the 25 phishing prevalence condition participants took significantly longer to classify phishing emails (~16 seconds) compared to legitimate emails (~13 seconds), $F(1,17) = 8.66$, $p = .009$, $\eta_p^2 = .34$. There were no differences in classifications times for phishing and legitimate emails for the other two conditions (p's > .414) (see Figure 8B). Similar to the accuracy results, additional analyses were conducted on response times for the same 5 phishing emails that each group received. This analysis determined that the number of total phishing emails influenced responses times for those same 5 phishing emails (see Figure 9B), $F(2,51) = 5.68$, $p = .006$, $\eta_p^2 = .25$. Pairwise comparisons revealed that this effect was largely driven by the 25 phishing prevalence condition taking longer (~24 seconds) than the 5 phishing prevalence condition (~13 seconds, $p = .003$) and the 50 phishing prevalence condition (~ 15 seconds, $p = .012$). There was no difference between the 50 phishing prevalence condition and the 5 phishing prevalence condition ($p = .575$) (see Figure 9B). Overall, these results indicate that under certain phishing prevalence rates, email users may take longer to evaluate phishing emails compared to regular emails.

A. B.

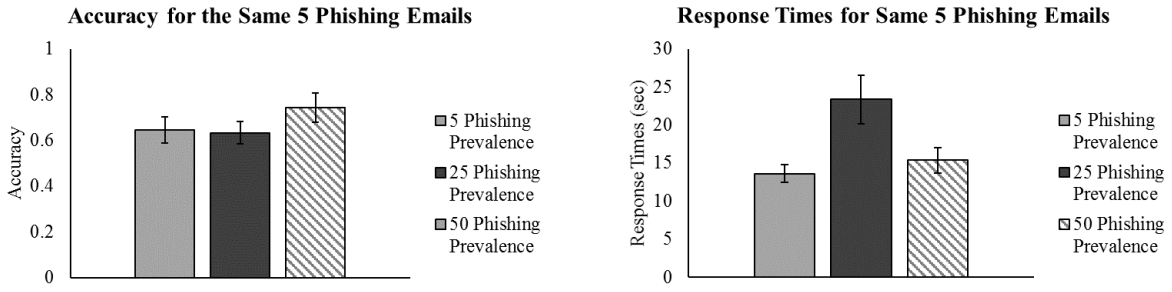


Figure 9. Experiment 2 Email Classification Accuracy (A) and Email Classification Response Times (B) for the Same 5 Phishing Emails by Phishing Prevalence

Error bars represent the standard error of the mean.

Sensitivity and Response Criterions

Just as with Experiment 1, signal detection measures were analyzed to better characterize phishing susceptibility. Both sensitivity and response criterions were calculated for each phishing prevalence group and subjected separately to two separate two-factor mixed ANOVAs with an alpha level of .05, with phishing prevalence (high, moderate, low) as the independent variable. Similar to Experiment 1, phishing prevalence was not predicted to influence sensitivity, but rather response criterions. Explicitly, participants who viewed fewer phishing emails were expected to be more liberal in their classifications and miss more phishing attacks.

A. B.

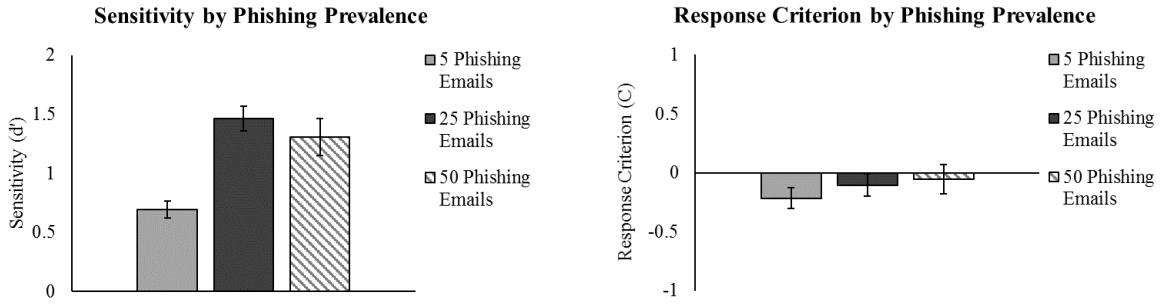


Figure 10. Experiment 2 Signal Detection Measures, Sensitivity (A) and Response Criterion (B) by Phishing Prevalence

Error bars represent the standard error of the mean.

Surprisingly, there was a main effect of phishing prevalence on sensitivity (see Figure 10A) , $F(2,51) = 8.42$, $p = .001$, $\eta_p^2 = .25$. Pairwise comparisons determined that this difference was based on the 5 phishing prevalence condition compared to the other two prevalence conditions. Specifically, that the lowest phishing prevalence condition had significantly lower sensitivity (0.69) compared to the moderate phishing prevalence condition (1.46, $p < .001$) and the high phishing prevalence condition (1.31, $p = .003$); the moderate and high phishing prevalence groups did not differ significantly from one another ($p = .437$). Interestingly phishing prevalence did not appear to impact response criterion (c) (see Figure 10B), $F(2,51) = 0.62$, $p = .545$, $\eta_p^2 = .02$. In order to determine if response criterions were considered liberal, separate one sample t-tests were conducted on each prevalence condition. The 5 phishing prevalence condition was the only condition significantly different than zero ($p = .023$), the other two groups were not (p's > .276). This suggests that only the 5 prevalence group was liberal in their responses, although they were not significantly different from the other two groups. Taken

together these results suggest that lowering the prevalence of phishing emails decreases email users' abilities to detect phishing emails without changing their response criterion.

Actions Chosen

The actions chosen for each email were also analyzed in a similar way to Experiment 1. Only phishing emails were analyzed, and "correct" actions consisted of checking the sender, deleting or reporting emails as suspicious, whereas "incorrect" actions consisted of clicking a link/opening an attachment or replying. Action choice accuracy was then submitted to a one-way between subjects' ANOVA with an alpha level of .05 with prevalence (high, moderate, low) as the independent variable. More incorrect actions were predicted for the low prevalence condition if decreasing the probability of a phishing email influenced behavior.

A. B.

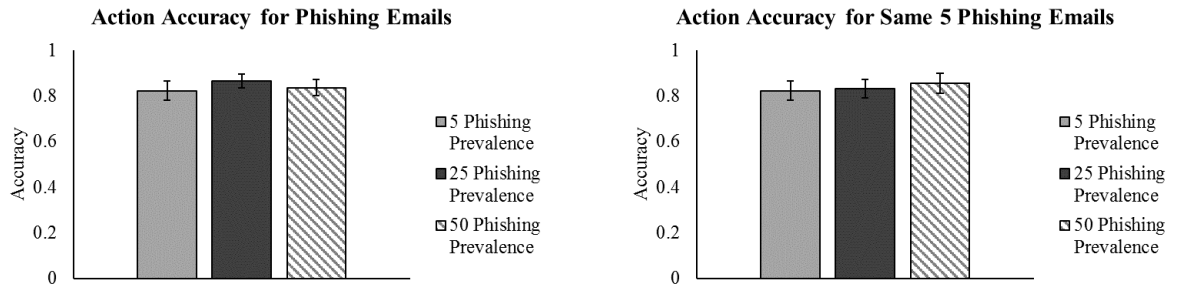


Figure 11. Experiment 2 Action Accuracy for Phishing Emails (A) and Action Accuracy for the Same 5 Phishing Emails (B)

Error bars represent the standard error of the mean.

There was a not a main effect of prevalence on action accuracy for phishing emails, $F(2,51) = 0.35$, $p = .705$, $\eta_p^2 = .01$, suggesting that prevalence does not change the actions selected for phishing emails (see Figure 11A) . Additionally, like the accuracy and response time

data, additional analyses were conducted for the same 5 phishing emails everyone received. These results indicated that once again phishing prevalence did not influence the actions selected (see Figure 11B) , $F(2,51) = 0.16$, $p = .854$, $\eta_p^2 = .01$. Overall, these results suggest that the prevalence of phishing emails does not influence the next action selected for emails.

Threat Level, Confidence, and Difficulty

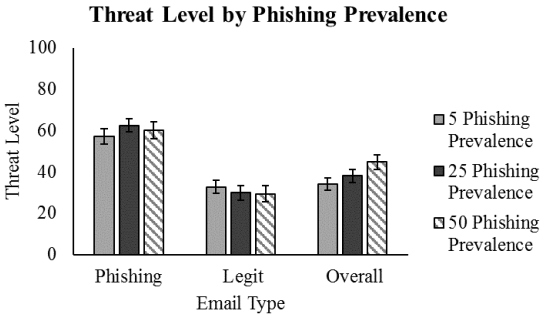
Threat level, confidence and difficulty were examined in the same way as Experiment 1 except they were explored in the context of the phishing prevalence. In order to determine if phishing prevalence influenced the three factors of threat level, confidence and difficulty, three separate two-factor mixed ANOVAs with alpha levels of .05, and prevalence (high, moderate, low) and email type (legitimate, phishing) as the independent variables were performed. Each measures' scores were calculated on across both correct and incorrect trials. If email users are less likely to detect phishing emails under conditions of low prevalence, then perceived threat level was expected to decrease under lower prevalence phishing conditions but confidence and difficulty were expected to stay the same.

Threat Level

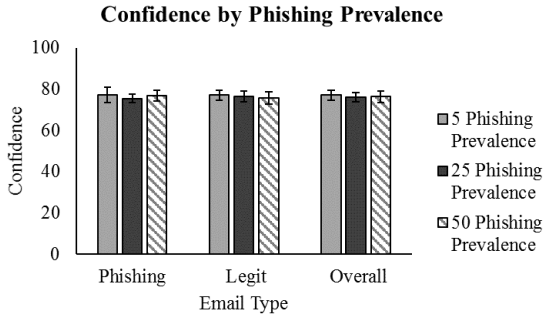
There was a main effect of email type on threat level, $F(1,51) = 248.61$, $p < .001$, $\eta_p^2 = .83$, such that phishing emails were rated as significantly more threatening (59.86) than legitimate emails (30.66) across all groups (see Figure 12A). There was not a main effect of phishing prevalence $F(2,51) = 0.06$, $p = .945$, $\eta_p^2 < .01$, or an interaction between email type and phishing prevalence, $F(2,51) = 1.77$, $p = .180$, $\eta_p^2 = .07$. There was a significant positive relationship between cyber experience and the perceived threat level of legitimate emails, r(54) =

.304, p = .025, indicating that more experience heightens the perceived threat of legit emails. These results suggest that prevalence does not change the perceived threat level of emails, and that threat level is solely determined by the legitimacy of the email.

A.



B.



C.

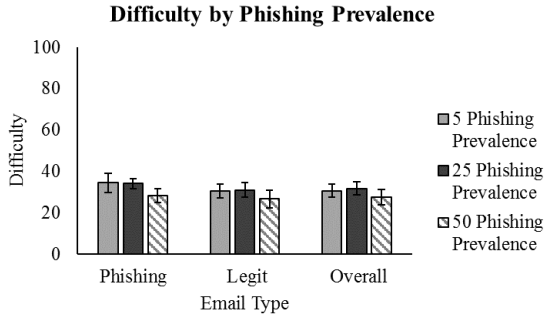


Figure 12. Experiment 2 Threat Level (A), Confidence (B) and Difficulty (C) ratings by Phishing Prevalence

Error bars represent the standard error of the mean.

Confidence

There were no main effects of email type, $F(1,51) = 0.01$, p = .919, $\eta_p^2 < .01$, or of phishing prevalence (see Figure 12B), $F(2,51) = 0.06$, p = .944, $\eta_p^2 < .01$, nor any interaction on

confidence, $F(2,51) = 0.25$, $p = .777$, $\eta_p^2 = .01$. These null effects suggest that confidence is unaffected by email type or phishing prevalence.

Difficulty

As with confidence, there were no main effects of email type, $F(1,51) = 2.34$, $p = .132$, $\eta_p^2 = .04$, or prevalence, $F(2,51) = 0.81$, $p = .457$, $\eta_p^2 = .03$, nor any interaction on perceived difficulty, $F(2,51) = 0.17$, $p = .845$, $\eta_p^2 = .01$ (see Figure 12C). Cyber experience was related to how difficult it was to evaluate phishing emails, $r(54) = .312$, $p = .022$, suggesting the more previous cyber experience participants had the more challenging they felt it was for them to evaluate phishing emails. This may simply be due to the fact that those who have enough cyber experience are more aware of the difficulty of this type of task. Overall, like confidence, difficulty level does not appear to be impacted by either phishing prevalence or email type.

CHAPTER FOUR: EXPERIMENT 3

Experiments 1 and 2 investigated how email task factors influenced phishing email detection. However, email load and phishing prevalence do not vary in isolation from one another in the real world. Thus, Experiment 3 examined the interaction of prevalence and email load. Sawyer et al., (2014) investigated a similar paradigm for an IP monitoring task when they manipulated event rates and the probability of a signal. Although this task is different from the email task at hand, the results can inform predictions for the current study. Sawyer et al., (2014) found that performance was poorest for conditions with the fast event rate, and low probability of a signal. This finding is consistent with the results found in Experiments 2 exploring the impact of prevalence but inconsistent with the null email load results found in Experiment 1. However, the influence of these task factors may differ when manipulated together. Therefore, it was predicted that participants would perform the worst when the prevalence rate of emails was the lowest, and the email load was the highest.

Method

Participants

Seventy-two participants (M_{age} = 18.45, 30 males, 42 females) were recruited from the University of Central Florida for course credit. All participants had normal or corrected-to-normal vision (20/32 or better corrected vision on a Snellen eye chart) and color vision (Ishihara's test for color blindness; 13 plates).

In order to determine how many participants were necessary to find an effect of the interaction of prevalence and email load a power analysis was conducted in G*Power (Faul et al., 2007). Sawyer et al., (2014) found an effect size of η_p^2 = .16, for the interaction of signal

probability and event rate. Both covariates from Experiment 1 (i.e., deficient self-regulation and cyber hygiene) were included, as well as the covariate from Experiment 2 (i.e., cyber experience). Therefore, an ANCOVA power analysis was conducted using a Cohen's *f* of .44, power of 0.95, an alpha probability of 0.01, 4 groups and 3 covariates. Based off this analysis, 72 participants (18 per group) should be sufficient to find a small effect size exploring the interaction of email load and phishing prevalence.

Apparatus and Stimuli

The apparatus and stimuli were the same as Experiment 1 with the following exceptions. As in Experiment 1, all participants evaluated 100 emails but their perceived email load was manipulated. In the low email load condition, they were told they had 100 emails in their inbox, and in the high email load condition they were told they had 300 emails in their inbox (see Figure 1). The number of phishing emails also varied based off the condition, with either low or high prevalence. The low prevalence condition contained 5% phishing emails, and the high prevalence condition contained 50% phishing emails. Lastly, participants were all given an hour timer that they could view throughout the experiment. This timer was implemented in order to enhance the email load effects from Experiment 1.

Individual Difference Measures

Experiment 3 utilized the same deficient self-regulation and cyber hygiene measures from Experiment 1 and the same cyber experience scale from Experiment 2.

Design

Design of Conditions

The number of phishing emails and the email load depended on the participants' assigned condition. The overall experimental design for Experiment 3 was a 2 (email load: high vs low) x 2 (prevalence: high vs low) x 2 (email type: phishing vs legitimate) mixed factorial design with the first two factors being between-subjects and the third factor being within-subjects. Thus, there were a total of four experimental groups that explored the interaction of phishing prevalence and email load. For the first two groups performance was examined under conditions of high perceived email load (300 emails in inbox) for both high phishing prevalence (50%, 50 emails) and low phishing prevalence (5%, 5 emails). The other two groups demonstrated phishing susceptibility under conditions of low perceived email load (100 emails in inbox) for both high phishing prevalence (50%, 50 emails) and low phishing prevalence (5%, 5 emails).

Design of Trials

The design of trials was the same as the trials in Experiment 1.

Procedure

The procedure for the current experiment was the same as Experiment 1 with the following exceptions. In addition to measuring deficient self-regulation and cyber hygiene, Experiment 3 included the previous cyber experience measure from Experiment 2. Additionally, participants were given an hour timer at the beginning of the email classification task. Participants were told that they only had an hour to classify all the emails and to alert their

experimenter if this timer ran out. If the timer ran out participants were told to keep going, which only happened for a small number of participants.

Results and Discussion

Similar to the first two experiments, data analysis was based primarily on accuracy and response times for the email classifications. Additional analyses were conducted on other aspects of the classifications including the threat level of the email, what action would be taken next, how difficult was it to assess, and how confident participants were in their responses.

Email Classifications

As with the first two experiments, the main analysis investigated if participants varied in their classification accuracy based off email prevalence and email load. The covariates were not related to any of the dependent variables except difficulty ratings, and therefore were only included in those analyses. Classification accuracy and response times were submitted to three-factor mixed ANOVAs with an alpha level of .05, and email load (high, low), phishing prevalence (high, low) and email type (legitimate, phishing) as the independent variables. Response times were calculated across both correct and incorrect trials. If realistic task factors like email load and email prevalence affect phishing detection, participants should have the most incorrect and quickest classifications when email load is high and email prevalence is low. Additionally, these factors may interact such that their influence may have compounding effects when both high email load and low prevalence are present, compared to when only one is present.

Email Classification Accuracy

There was not a main effect of email type on accuracy, $F(1,68) = 0.83$, $p = .366$, $\eta_p^2 = .01$, suggesting that participants did not differ in their ability to classify legitimate and phishing emails (see Figure 13A). There was no interaction between email type and phishing prevalence/email load (p's >.163). There was also not a main effect of email load, $F(1,68) = 0.01$, $p = .911$, $\eta_p^2 < .01$, or phishing prevalence, $F(1,68) = 0.02$, $p = .897$, $\eta_p^2 < .01$. However, there was an interaction of email load and prevalence, $F(1,68) = 5.51$, $p = .002$, $\eta_p^2 = .08$. In order to break this interaction down separate ANOVAs were conducted on each prevalence condition. When there were only five phishing emails present, there was no effect of email load, $F(1,34) = 0.1.89$, $p = .178$, $\eta_p^2 = .05$. However, when there were 50 phishing emails present there was a difference between the two email load conditions, $F(1,34) = 4.45$, $p = .042$, $\eta_p^2 = .12$, such that the 100 email load condition had higher accuracy (~71%) compared to the 300 email load condition (~66%). Overall these results suggest that email load is an important predictor of classification accuracy, at least under circumstances where there is a 50/50 split between legitimate and phishing emails.

A. B.

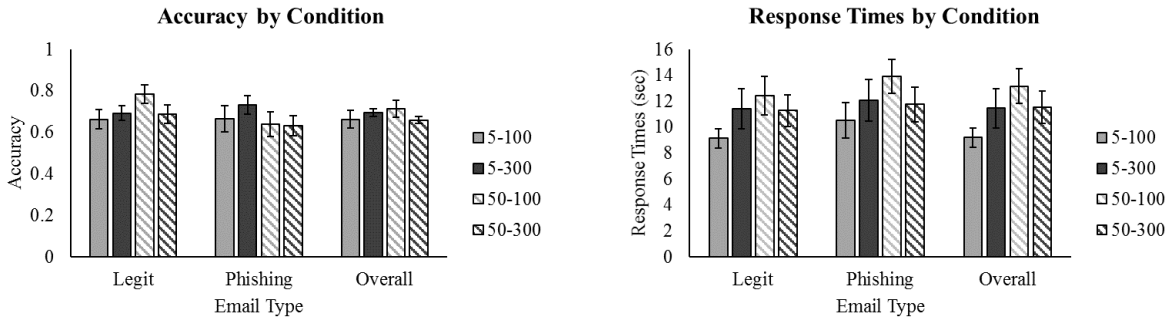


Figure 13. Experiment 3 Email Classification Accuracy (A) and Email Classification Response Times (B) by Email Load and Phishing Prevalence

Error bars represent the standard error of the mean.

Like Experiment 2, the same 5 phishing emails were also analyzed in order to examine a more direct comparison between the prevalence conditions (see Figure 14A). There was no significant difference between the email load conditions, $F(1,68) = 1.34$, $p = .250$, $\eta_p^2 = .02$, or the interaction between email load and phishing prevalence, $F(1,68) = 0.34$, $p = .564$, $\eta_p^2 < .01$. There was a marginal difference between the phishing prevalence conditions, $F(1,68) = 3.73$, $p = .058$, $\eta_p^2 = .05$, such that participants in the 5 phishing prevalence condition were more accurate (~72%) than the 50 phishing prevalence condition (~61%). Although surprising, this result indicates that the lower phishing prevalence condition may have helped participants classify the few phishing emails they viewed relative to the higher phishing prevalence condition.

A. B.

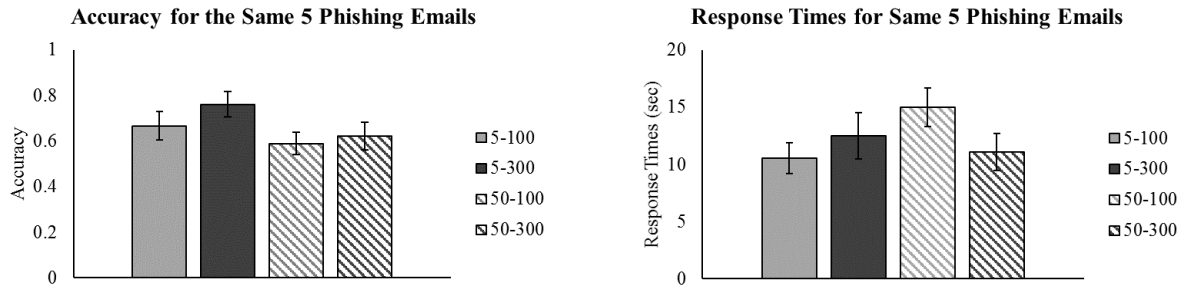


Figure 14. Experiment 3 Email Classification Accuracy (A) and Email Classification Response Times (B) for the Same 5 Emails by Email Load and Phishing Prevalence

Error bars represent the standard error of the mean.

Email Classification Response Times

There was a main effect of email type on response times, $F(1,68) = 5.57$, $p = .021$, $\eta_p^2 = .08$, such that participants took longer to evaluate phishing emails (~12 seconds) than legitimate emails (~11 seconds) (see Figure 13B). There was not a main effect of prevalence, $F(1,68) = 1.46$, $p = .231$, $\eta_p^2 = .02$, or email load, $F(1,68) = 0.01$, $p = .918$, $\eta_p^2 < .01$. There were no significant interactions (p's >167). Like the accuracy analyses, response times were examined for the same 5 phishing emails that all participants classified (see Figure 14B). There were still no main effects of email load, $F(1,68) = 0.18$, $p = .672$, $\eta_p^2 < .01$, and prevalence, $F(1,68) = 0.56$, $p = .458$, $\eta_p^2 = .01$. The interaction of email load and prevalence trended toward, but did not reach significance, $F(1,68) = 3.51$, $p = .065$, $\eta_p^2 = .05$. Lastly, there was a significant relationship between cyber hygiene and phishing response times, $r(72) = .233$, $p = .049$, suggesting that the more cyber hygiene participants reported the longer it took them to evaluate phishing emails. Overall these results indicate that the main factor contributing to differences in classification times in our task is the legitimacy of the email and the participant's cyber hygiene.

Sensitivity and Response Criterions

In a similar vein to the first two experiments, signal detection measures were analyzed to more fully understand phishing susceptibility under varying email load and phishing email prevalence. Both sensitivity and response criterions were calculated for each group. Sensitivity and response criterions were subjected to two separate three-factor mixed ANOVAs with an alpha level of .05, with email load (high, low), and phishing prevalence (high, low) as the independent variables. As with the first two experiments phishing prevalence and email load were not predicted to influence sensitivity but instead influence response criterions. Specifically, participants who saw fewer phishing emails under high levels of email load were expected to be more liberal in their classifications (i.e., rate more emails as phishing) and miss more phishing attacks.

There was no main effect of email load on sensitivity measures, $F(1,68) = 0.52$, $p = .474$, $\eta_p^2 = .01$, such that individuals did not differ in their sensitivities by email load (see Figure 15A). There was not a main effect of prevalence on sensitivity, $F(1,68) = 0.70$, $p = .407$, $\eta_p^2 = .01$, nor an interaction between prevalence and email load, $F(1,68) = 0.57$, $p = .452$, $\eta_p^2 = .01$. It is important to note that, like Experiments 1 and 2, these sensitivities are extremely low and all individuals were very poor at this task. Overall, email load and phishing prevalence do not seem to negatively impact phishing sensitivity and all users struggle to classify emails.

A. B.

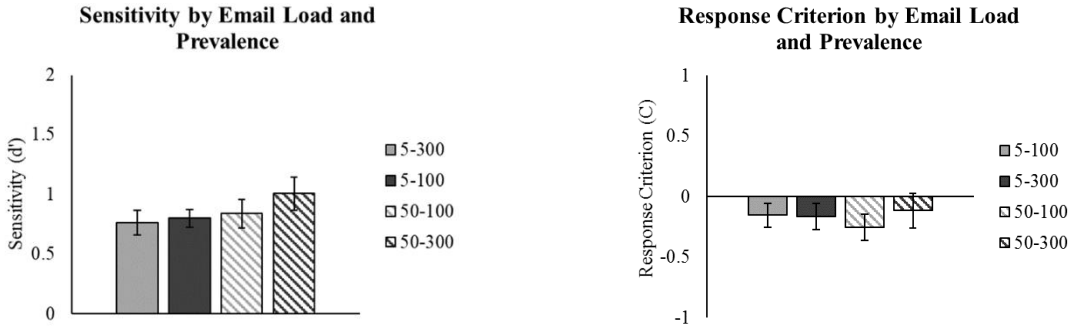


Figure 15. Experiment 3 Signal Detection Measures, Sensitivity (A) and Response Criterion (B) by Email load and Phishing prevalence

Error bars represent the standard error of the mean.

There were no main effects of email load, $F(1,68) = 0.29$, $p = .294$, $\eta_p^2 < .01$, or phishing prevalence, $F(1,68) = 0.05$, $p = .828$, $\eta_p^2 < .01$, nor any interaction on response criterion, $F(1,68) = 0.38$, $p = .538$, $\eta_p^2 = .01$ (see Figure 15B). Like the first two experiments each groups' response criterion was submitted to separate one-sample t-tests to determine if they were significantly different from zero. Only the 50 prevalence, 100 email load condition was determined to be significantly different from zero ($p = .033$) suggesting they were liberal in their responses. The other groups were not different from zero (p's >.130) suggesting that they were unbiased. Overall, the effects of email load and phishing prevalence did not appear to influence response criterion.

Actions Chosen

The actions chosen for each email were again analyzed. Identical to Experiments 1 and 2, only phishing emails were included in the analyses exploring actions, with correct action choices including checking the sender, deleting the email or reporting emails as suspicious, and incorrect

action choices including clicking a link/opening an attachment or replying. Action choice accuracy was then submitted to a one-way between subjects ANOVA with an alpha level of .05 with email load (high, low) and phishing prevalence (high, low) as the independent variables. As with the first two experiments, if high email load and low phishing prevalence decrease phishing detection, then riskier (i.e., more incorrect) actions were expected.

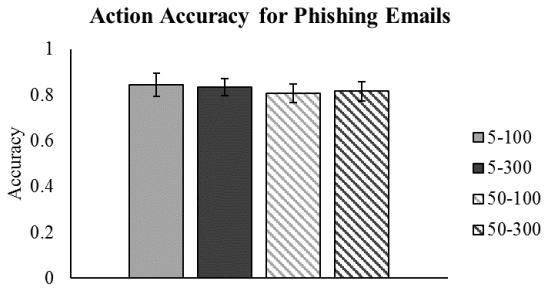


Figure 16. Experiment 3 Action Accuracy for Phishing Emails

Error bars represent the standard error of the mean.

There were no main effects of email load, $F(1,68) = 2.14$, $p = .148$, $\eta_p^2 = .03$, or phishing prevalence, $F(1,68) = 0.06$, $p = .803$, $\eta_p^2 < .01$, nor their interaction on action accuracy, $F(1,68) = 0.35$, $p = .555$, $\eta_p^2 < .01$(see Figure 16). These results suggest that neither email load or phishing prevalence, nor the interaction of these two factors influence email actions.

Threat Level, Confidence, and Difficulty

Like the first two experiments, threat level, confidence, and difficulty were explored in the context of email load and phishing prevalence and calculated across both correct and incorrect trials. Two separate three-factor mixed ANOVAs with alpha levels of .05, with email load (high, low), prevalence (high, low) and email type (legitimate, phishing) as the independent

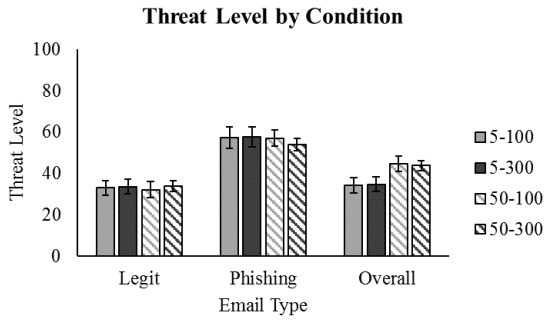
variables were performed to investigate the relationship between email load and prevalence on threat level and confidence. As cyber hygiene was related to the difficulty ratings for both legitimate, r(70) = -.24, *p* =.039, and phishing emails, r(70) = -.24, *p* =.046, it was included as a covariate. Additionally, cyber experience was related to both legitimate, r(70) = -.25, *p* =.035, and phishing email difficulty ratings, r(70) = .25, *p* =.036, so it was also included as a covariate. Thus, difficulty ratings were submitted to a three-factor mixed ANCOVA with alpha levels of .05, with email load (high, low), prevalence (high, low) and email type (legitimate, phishing) as the independent variables, and cyber hygiene and experience as covariates. It was expected that perceived threat level would be lowest when email load was high and phishing prevalence was low, and highest when email load was low and phishing prevalence was high. These relationships were predicted if threat level varies in a similar way to the email legitimacy classification. Confidence was expected to also vary by email load and prevalence such that confidence would be lowest under conditions of high email load and high prevalence, and highest under low email load and low prevalence. Confidence was predicted to vary in this manner if participants exhibited a false sense of confidence under the poorest task conditions. Lastly, difficulty was hypothesized to be the lowest under conditions of low email load and low phishing prevalence, and the highest under high email load and high phishing prevalence. These relationships were expected if users are aware of the increasing task demand of higher email loads but unaware of their decreased detection ability in lower prevalence conditions.

Threat Level

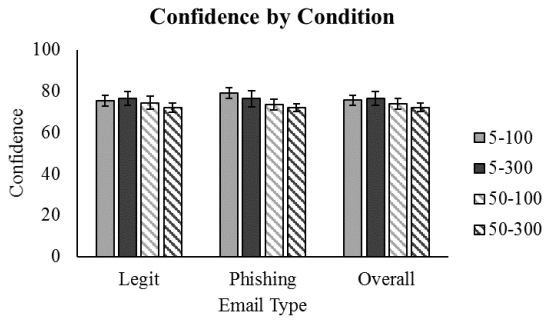
There was a main effect of email type on threat level, $F(1,68) = 250.53$, $p < .001$, $\eta_p^2 = .78$, such that participants rated phishing emails as more threatening (56.59), than legitimate

emails (33.12) (see Figure 17A). There was not a main effect of phishing prevalence, $F(1,68) = 0.11$, $p = .741$, $\eta_p^2 < .01$, nor email load on threat level, $F(1,68) < 0.01$, $p = .965$, $\eta_p^2 < .01$. There were also no significant interactions (p's >.387). Overall, these results indicate that level of threat perceived for emails depends solely on the legitimacy of the email rather than other task factors.

A.



B.



C.

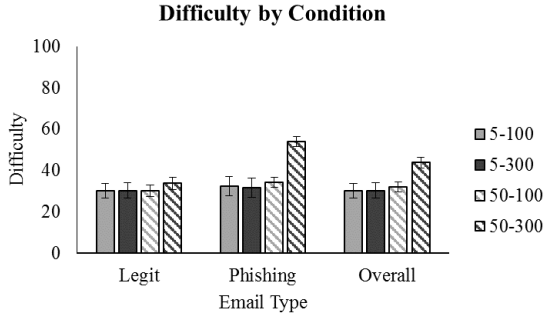


Figure 17. Experiment 3 Threat Level (A), Confidence (B) and Difficulty (C) ratings by Phishing Prevalence

Error bars represent the standard error of the mean.

Confidence

There were no main effects of email type, $F(1,68) = 0.42$, $p = .518$, $\eta_p^2 < .01$, email load, $F(1,68) = 0.23$, $p = .636$, $\eta_p^2 < .01$, or phishing prevalence on confidence, $F(1,68) = 2.07$, $p = .154$, $\eta_p^2 = .03$. There were also no significant interactions (p's >.224) (see Figure 17B). These results suggest that confidence levels are ubiquitous regardless of the legitimacy of the email or various task factors.

Difficulty

Cyber hygiene was found to be a significant covariate for difficulty ratings, $F(1,66) = 5.09$, $p = .027$, $\eta_p^2 = .07$. Cyber experience was also a significant covariate, $F(1,66) = 4.70$, $p = .034$, $\eta_p^2 = .07$. However, after controlling for both cyber hygiene and cyber experience there was not a main effect of email type, $F(1,66) = 0.51$, $p = .479$, $\eta_p^2 = .01$, email load, $F(1,66) = 0.36$, $p = .553$, $\eta_p^2 = .01$, or phishing prevalence, $F(1,66) = 0.49$, $p = .488$, $\eta_p^2 = .01$ (see Figure 17C). Additionally, there were no significant interactions (p's >.646). Lastly, there was a significant relationship between impulsivity (i.e., BIS-11 scores) and difficulty ratings for legitimate emails, $r(72) = .286$, $p = .015$, indicating that the more impulsive participants were, the more challenging they found classifying legitimate emails. Overall, these results suggest that task factors and the legitimacy of emails do not dictate how difficult the task is, but rather individual difference factors such as cyber hygiene and experience.

CHAPTER FIVE: EXPERIMENT 4

The first three experiments identified situations in which email users are the most susceptible to phishing attacks. However, there has been limited success in the previous research aimed at improving phishing detection in these situations. The West Point Carronade (Ferguson, 2005) demonstrated how even trained cadets fell victim to phishing attacks 90% of the time. Attempts at training individuals to be resilient to phishing attacks have either been specific to the email testbed (Sawyer et al., 2015), resulted in poor overall accuracy (Kumaraguru et al., 2007b), or were limited in retention (Mayhorn & Nyeste, 2012). Experiment 4 took a novel approach to improve phishing detection by investigating interventions rather than training. Byrne et al., (2016) suggested that a cheat sheet of information may prove useful to improve cyber performance. Experiment 4 utilized a cheat sheet that included information regarding the typical characteristics of phishing emails identified by previous research (Bergholz et al., 2010; Chandrasekaran, Narayanan, & Upadhyaya, 2006; Drake et al., 2004) Additionally, Kumaraguru et al., (2007b) demonstrated that embedded training elicits better performance compared to non-embedded methods that utilize the same training information (e.g., cartoons, pamphlets). Thus, Experiment 4 included a cheat sheet with the same information identified above, but embedded into the task. In order to understand how the cheat sheet benefits phishing detection, Experiment 4 also implemented a control condition in which participants received no intervention and completed the task normally. Based on previous research, it was expected that the physical cheat sheet would improve performance relative to the control (Byrne et al., 2016), but the best phishing detection was expected to occur in the embedded cheat sheet condition. Additionally, as research has shown that conscientiousness and agreeableness can influence how likely an

individual is to utilize cyber interventions (McBride et al., 2012; Shropshire et al., 2006; Shropshire et al., 2015), it was expected that individuals who rated highly on both characteristics would detect more phishing emails in both intervention methods relative to the control group because they would have utilized the cheat sheet information more.

Method

Participants

Fifty-seven participants (M_{age} = 16.35, 19 males, 38 females) from the University of Central Florida participated in this study in exchange for course credit. All participants had normal or corrected-to-normal vision (20/32 or better corrected vision on a Snellen eye chart) and color vision (Ishihara's test for color blindness; 13 plates).

A power analysis was conducted in G*Power (Faul et al., 2007) to ensure that Experiment 4 would have enough participants to find an effect of training intervention. Sawyer et al., (2105) found an effect size of η_p^2 = .23, for the impact of training on email performance. Additionally, previous research has shown that various personality factors can influence the utilization of cyber interventions. Specifically, that individuals who rate high on conscientiousness and agreeableness are more likely to utilize interventions and follow cyber protocols (McBride et al., 2012; Shropshire et al., 2006; Shropshire et al., 2015). Thus, an ANCOVA power analysis was calculated using a Cohen's *f* of .55, power of 0.95, an alpha probability of 0.01, 3 groups and 2 covariates. Based off of this analysis, 57 participants (19 per group) should be sufficient to find a small effect size exploring the impacts of various intervention types.

Apparatus and Stimuli

The apparatus and stimuli were identical to Experiment 1 with the following exceptions. In order to explore how cyber interventions improve performance for phishing detection, Experiment 4 utilized the poorest task conditions (i.e., high email load & low phishing prevalence). Meaning, all participants were deceived into believing they needed to evaluate 300 emails, when in reality they only viewed 100 emails, with 5% (i.e., 5 emails) being phishing attempts. Additionally, participants who received an intervention either saw a cheat sheet regarding tips for detecting phishing emails (see Figure 18) or the same information embedded in the GMAIL interface (see Figure 19). This information consisted of the typical characteristics of phishing emails posed as questions. For instance, “does the email have a plausible premise?”. A literature review first identified twenty characteristics that are typical of phishing emails. A pilot study conducted by Sarno, Lewis, Shoss, Bohil and Neider then narrowed down the characteristics to seven that are the most predictive of a phishing email in the present email set. The five remaining characteristics included an implausible premise (Bergholz et al., 2010; Drake et al., 2004), time pressure (Drake et al., 2004), collecting personal information (Drake et al., 2004), account deletion/suspension threats (Chandrasekaran et al., 2006), and spelling or grammatical errors. The last characteristic was identified as a consistent theme in the present email set by researchers.

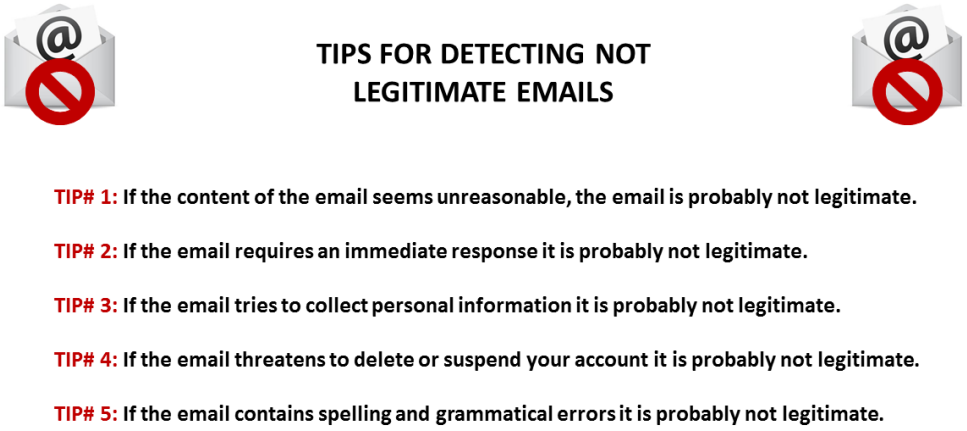


Figure 18. Phishing Email Non-Embedded Cheat Sheet

Participants in this condition were given a cheat sheet of information to assist in their classification of emails. The cheat sheet included tips that indicated qualities of phishing (or not legitimate) emails (e.g., collecting personal information).

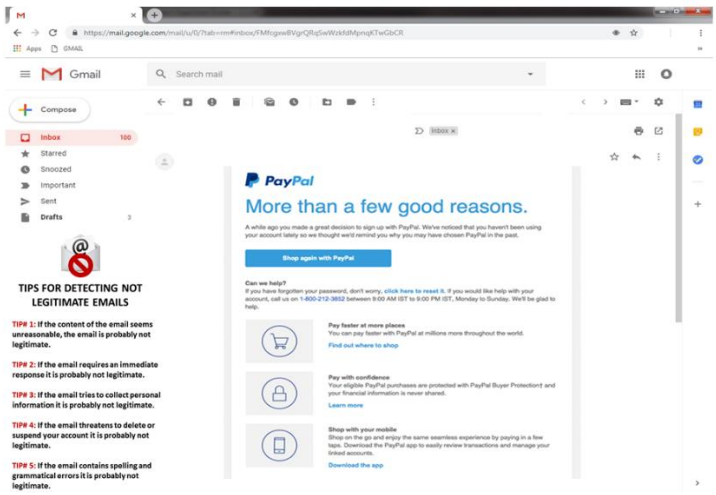


Figure 19. Phishing Email Embedded Cheat Sheet

Participants in this condition were given the same information from the non-embedded physical cheat sheet condition but embedded into the GMAIL interface.

Individual Difference Measures

The Big Five Inventory Modified: Conscientiousness

The Big Five Inventory Modified: Conscientiousness (John, Donahue, & Kentle, 1991) is a 9-item scale measuring the personality characteristic of conscientiousness. Items are rated on a 5-point Likert-type scale with anchors ranging from disagree strongly to agree strongly. Example items include, “is a reliable worker” and “does things efficiently”. After reverse scoring, high scores indicated more conscientiousness.

The Big Five Inventory Modified: Agreeableness

The Big Five Inventory Modified: Agreeableness (John et al., 1991) is a 9-item scale measuring the personality characteristic of agreeableness. Items are rated on a 5-point Likert-type scale with anchors ranging from disagree strongly to agree strongly. Example items include, “is helpful and unselfish with others” and “is generally trusting”. After reverse scoring, high scores indicated more agreeableness.

Design

Design of Conditions

Participants were assigned to one of three cyber interventions. The first condition was a physical cheat sheet condition where participants were provided with a printout version of the information provided in Figure 18. The other intervention condition included the same information, but was provided to participants embedded in the task in the GMAIL interface (see Figure 19). Participants were allowed to view the embedded information while evaluating the

emails, so it was comparable to the physical cheat sheet condition. There was also a control condition where participants proceeded through the task without any intervention. The overall design for Experiment 4 was a 3 (intervention type: none vs physical cheat sheet vs embedded cheat sheet) by 2 (email type: legitimate vs not legitimate) mixed design with the first factor being between-subjects and the second factor being within-subjects.

Design of Trials

The design of trials was the same as that of Experiment 1.

Procedure

The procedure utilized in Experiment 4 was the same as in Experiment 1 with the following differences. Instead of participants' deficient self-regulation and cyber hygiene being measured, participants completed a modified Big Five inventory (John et al., 1991) to determine how they rated on conscientiousness and agreeableness. Participants in the intervention conditions also received instructions regarding how to utilize the cheat sheet. Like Experiment 3, participants were provided with a timer to keep track of how much time they had left to classify emails.

Results and Discussion

As with the first three experiments, data analysis was based primarily on accuracy and response times for the email classifications. Additional analyses were conducted on other aspects of the classifications including the threat level of the email, what action would be taken next, how difficult was it to assess, and how confident participants were in their responses.

Email Classifications

The primary goal of Experiment 4 was to determine which intervention method best improved performance relative to the control condition. Neither conscientiousness or agreeableness was correlated with any of the dependent variables and were therefore removed from any further analyses. Thus, classification accuracy and response times were each submitted to a two-factor mixed ANOVA with an alpha level of .05, with intervention type (none, physical cheat sheet, embedded cheat sheet) and email type (legitimate, phishing) as the independent variables. Response times were calculated across both correct and incorrect trials. The best intervention method should result in the highest classification accuracy and longest classification times relative to the control group.

Email Classification Accuracy

There was not a main effect of email type on accuracy, $F(1,54) = 1.43$, $p = .238$, $\eta_p^2 = .03$, indicating that participants evaluated legitimate and phishing emails equally well (see Figure 20A). There was also not a main effect of intervention, $F(2,54) = 1.77$, $p = .181$, $\eta_p^2 = .06$, suggesting that our interventions did not differ from the control group. However, there was a significant interaction between condition and email type, $F(2,54) = 4.03$, $p = .023$, $\eta_p^2 = .13$. In order to explore this interaction additional ANOVA's were conducted on each email type (i.e., legitimate, and phishing) separately. There results revealed that there were no differences between the interventions for legitimate emails, $F(2,54) = 2.01$, $p = .143$, $\eta_p^2 = .07$, but there were differences between the groups for phishing emails, $F(2,54) = 3.96$, $p = .025$, $\eta_p^2 = .13$. Pairwise comparisons revealed that participants in the embedded cheat sheet condition detected more phishing emails (79%) than the control group (56%, $p = .008$). There was no difference

between the control group and the physical cheat sheet group (63%, p = .385). There was a trend towards a benefit of the embedded cheat sheet over the physical cheat sheet group (p = .066) but it was not significant. Overall, these results suggest that our embedded intervention assisted participants in detecting the low prevalence phishing emails relative to the control group.

A. B.

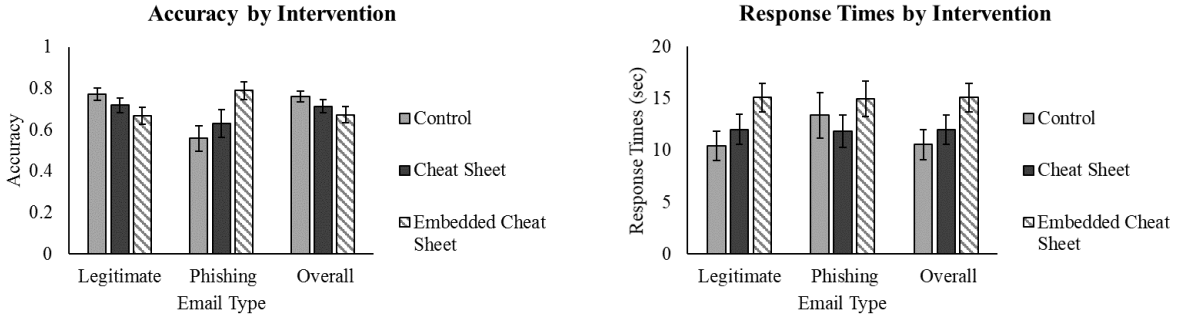


Figure 20. Experiment 4 Email Classification Accuracy (A) and Email Classification Response Times (B) by Intervention

Error bars represent the standard error of the mean.

Email Classification Response Times

There was no main effect of email type on response times, $F(1,54) = 2.39$, $p = .128$, $\eta_p^2 = .04$, suggesting that participants took the same amount of time to evaluate both phishing and legitimate emails (see Figure 20B). There was also not a main effect of intervention, $F(2,54) = 1.27$, $p = .290$, $\eta_p^2 = .05$, suggesting that our interventions did not change the amount of time participants spent evaluating emails. Similarly to the accuracy data, there was a significant interaction between email type and intervention, $F(2,54) = 3.24$, $p = .047$, $\eta_p^2 = .11$. Additional ANOVA's on each email type revealed that there were no differences between the intervention

groups (*p*'s >.082). These analyses might be underpowered, or the interaction may be spurious. Overall these results suggest that all participants took roughly the same amount of time to evaluate both types of emails regardless of intervention group.

Sensitivity and Response Criterions

Signal detection measures were again analyzed, but this time to compare benefits of the two interventions. Both sensitivity and response criterions were calculated for each intervention group and subjected separately to two separate one-way between subject's ANOVAs with an alpha level of .05, with intervention type (none, physical cheat sheet, embedded cheat sheet) as the independent variable. Unlike the first three experiments, if the interventions were truly improving performance, then sensitivity was expected to improve. Response criterion shifts were also expected. Specifically, that participants may only improve in accuracy because the interventions result in them being more conservative with their classifications, rating more emails as phishing.

Surprisingly, there were no main effects of intervention on sensitivity (d'), $F(2,54) = 1.00$, $p = .376$, $\eta_p^2 = .04$ (see Figure 21A), or response criterion (C), $F(2,54) = 2.80$, $p = .070$, $\eta_p^2 = .09$ (see Figure 21B). Each intervention conditions' response criterion was submitted to a separate one sample t-test to determine if they were significantly different from zero. Both the control condition's and the cheat sheet condition's response criterion were significantly different than zero ($p = .004$, $p = .006$, respectively), suggesting they were liberal in their responses. However, the embedded cheat sheet group was not different from zero ($p = .747$), suggesting

they were unbiased in their responses. Taken together these results suggest participants do not have systematically different sensitivities or response criterions due to the present interventions.

A. B.

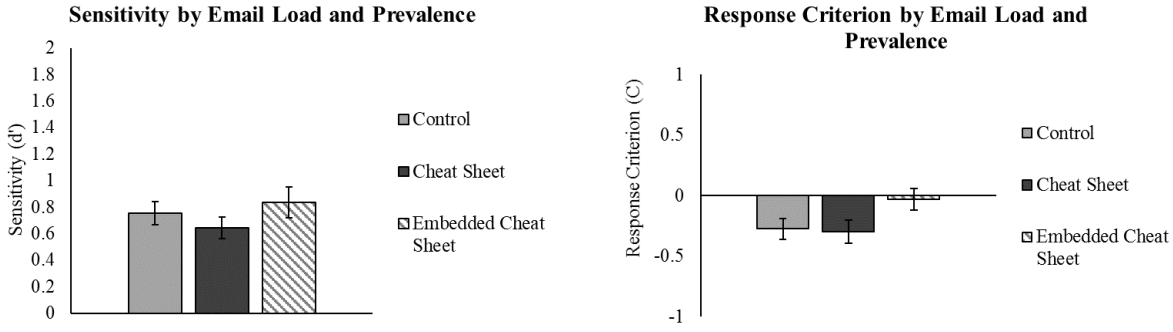


Figure 21. Experiment 4 Signal Detection Measures, Sensitivity (A) and Response Criterion (B) by Intervention

Error bars represent the standard error of the mean.

Actions Chosen

As with the first three experiments, action analyses were limited to phishing emails only. Specifically, accuracy was determined by whether participants selected an appropriate action for each email (e.g., delete) or not (e.g., reply). Accuracy for actions chosen were then submitted to a one-way between subjects ANOVA with an alpha level of .05, with intervention type (none, physical cheat sheet, embedded cheat sheet) as the independent variable. If intervention type improves the appropriate actions selected, then it was expected that the embedded cheat sheet would result in the highest action choice accuracy.

There was no main effect of intervention on action accuracy, $F(2,54) = 2.70$, $p = .076$, $\eta_p^2 = .09$, suggesting that although or interventions changed classification accuracy they did not change the actions selected (see Figure 22).

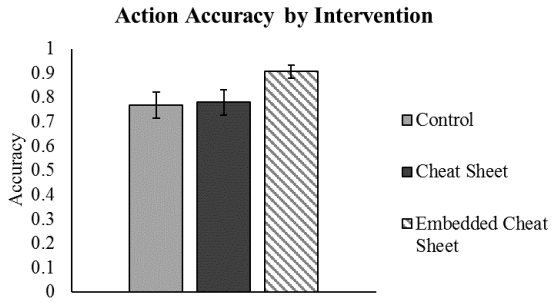


Figure 22. Experiment 4 Action Accuracy for Phishing Emails by Intervention

Error bars represent the standard error of the mean.

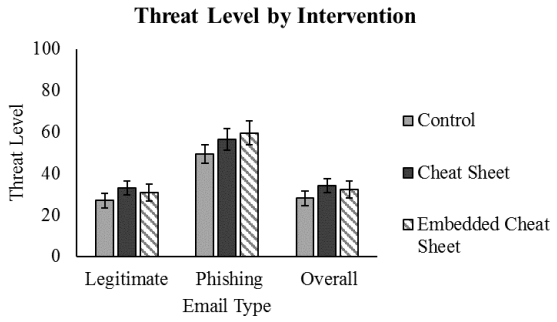
Threat Level, Confidence, Difficulty

If intervention type influences participants' classifications and action selections, then it was also expected to affect their perceived threat level, confidence, and difficulty. To examine these relationships, three separate two-factor mixed ANOVAs were conducted with alpha levels of .05, and intervention type (none, physical cheat sheet, embedded cheat sheet) and email type (legitimate, phishing) as the independent variables. All measures were calculated across both correct and incorrect responses. The embedded cheat sheet was expected to improve performance the most, which should have resulted in increased confidence, increased perceived threat level, and decreased perceived difficulty. The physical cheat sheet was also expected to improve performance in these three factors relative to control. Understanding how the two intervention methods impact these other aspects of email classifications may assist in evaluating and improving them.

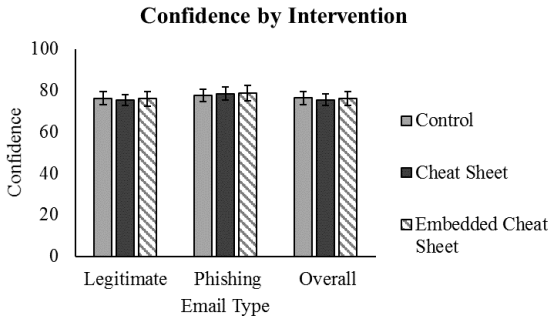
Threat Level

There was a significant main effect of email type on threat level, $F(1,54) = 121.58$, $p < .001$, $\eta_p^2 = .69$, such that participants rated legitimate emails as less threatening (30.26) than phishing emails (55.06) (see Figure 23A). There was not a main effect of intervention, $F(2,54) = 0.96$, $p = .390$, $\eta_p^2 = .03$, nor an interaction between intervention and email type, $F(2,54) = 0.71$, $p = .496$, $\eta_p^2 = .03$, suggesting that only the legitimacy of the email dictated the perceived threat level. Lastly, there was a significant relationship between agreeableness and the perceived threat levels of phishing emails, $r(57) = -.301$, $p = .023$, indicating that the more agreeable participants were, the less threatening they found the phishing emails.

A.



B.



C.

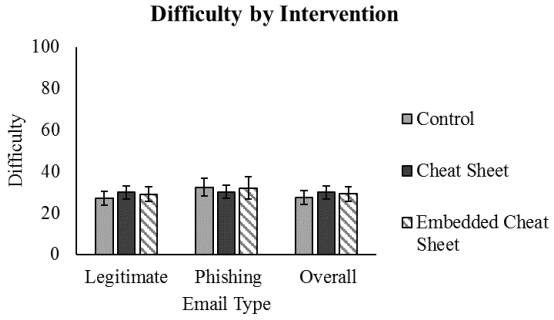


Figure 23. Experiment 4 Threat Level (A), Confidence (B) and Difficulty (C) ratings by Intervention

Error bars represent the standard error of the mean.

Confidence

There was also a main effect of email type on confidence, $F(1,54) = 4.78$, $p = .033$, $\eta_p^2 = .08$, such that participants were slightly more confident for phishing emails (78.30) than legitimate emails (75.87) (see Figure 23B). There was not a main effect of intervention, $F(2,54) = 0.01$, $p = .992$, $\eta_p^2 < .01$, or an interaction between email type and intervention, $F(2,54) = 0.23$,

p = .797, η_p^2 =.01. Overall, these results suggest that like the threat level data, confidence ratings were largely driven by the email type not the intervention.

Difficulty

There were no main effects of email type, $F(1,54) = 2.02$, p = .161, η_p^2 =.04, intervention, $F(2,54) = 0.01$, p = .987, $\eta_p^2 < .01$, nor any interaction on difficulty, $F(2,54) = 0.59$, p = .587, η_p^2 =.02 (see Figure 23C). These results indicate that the difficulty of this task was not influenced by the type of email or the intervention utilized.

CHAPTER SIX: GENERAL DISCUSSION

Previous work that has explored cybersecurity performance in email contexts has suggested that email users are very poor at detecting phishing emails (Ferguson, 2005). Even studies that have implemented training methodologies have struggled at improving performance to a level that makes users sufficiently resilient to phishing attacks (Kumaraguru et al., 2007a; Kumaraguru et al., 2007b; Mayhorn & Nyeste, 2012; Sawyer et al., 2015). Additionally, few studies have examined how email load (Vishwanath et al., 2011) and the prevalence of phishing emails (Sawyer & Hancock, 2018) influence the detection of phishing emails, and no studies have examined how these two factors may interact.

The present experiments remedy the previously mentioned gaps in the literature by investigating the effects of both email load and phishing prevalence, and what intervention may be best suited to improve performance under such task conditions. Experiment 1 explored how high email load may negatively influence email classification for the first time in an experimental setting. Experiment 2 looked at how low prevalence settings decrease phishing detection with a novel, more diverse set of emails than previously utilized (Sawyer & Hancock, 2018). Experiment 3 utilized the two variables of the first two experiments (email load: high vs low, phishing prevalence: high vs low) to investigate if these task factors interact, thus creating even poorer performance under conditions of high email load and low phishing prevalence. Experiment 4 utilized the high email load and low phishing prevalence conditions to help develop an intervention aimed at improving correct phishing detection under the worst task factors. Lastly, all four experiments utilized several individual difference variables to help identify how various cognitive (i.e., deficient self-regulation), behavioral (i.e., previous cyber

experience, cyber hygiene) and personality factors (i.e., agreeableness, conscientiousness) influence phishing detection under varying email task conditions and the utilization of training interventions.

The Effect of Task Factors

Experiments 1 through 3 explored the effects of various task factors on email classifications. No previous research has manipulated how the number of emails in a user's inbox influences their ability to detect phishing emails. Email load was manipulated in both Experiment 1 and 3, and the results indicated that the more emails a user has in their inbox (e.g., 300 emails vs 100 emails) the more difficult it is to classify emails. Additionally, in Experiment 3 higher email loads decreased accuracy (at the 50% phishing prevalence condition). Taken together, these results indicate that having more emails in your inbox may negatively impact your ability to correctly classify emails. This is particularly important in the real world given that the average working professional has over 20 unread emails in their inbox and get 120 new emails every day (Plummer, 2019). These results suggest that email systems should implement restrictions on how many emails you are able to interact with at once to prevent high email load. This in turn may decrease the difficulty of the task and decrease vulnerability to phishing emails.

The prevalence of phishing emails was also manipulated in Experiments 2 and 3. In Experiment 2 lower phishing prevalence resulted in poorer overall accuracy and decreased sensitivity. These results are consistent with previous findings that demonstrate decreased phishing detection with fewer phishing emails (Sawyer et al., 2015). Surprisingly, in Experiment 3 the low prevalence condition had higher accuracy compared to the high prevalence condition for the same emails. These two experiments together suggest that there may be a complex

relationship between phishing prevalence and fraud detection. Specifically, that there may be some situations where more phishing emails make it more challenging to decipher legitimate emails from fraudulent ones. In a recent training study from Singh, Aggarwal, Rajivan, and Gonzalez (2019), higher phishing prevalence training conditions (75% phishing) decreased sensitivity for phishing detection following training. Minor sensitivity improvements were only seen when phishing prevalence rates were at 25% and 50%. Thus, email users seem to lack sensitivity for phishing emails (even with training) regardless of the number of phishing emails present and including more phishing emails only provided more opportunities to miss attacks. This is an important finding, because it suggests that users may not benefit from simulated phishing attacks, at least those that deviate from low prevalence rates.

Individual Differences in Email Classifications

Individual differences have been explored by several previous phishing studies (e.g., Sarno et al., in press; Sheng et al, 2011). However, there has been limited work exploring how these traits may influence the different aspects of email classifications. Although there were very limited relationships between the individual difference variables and the dependent measures in the current studies, some interesting patterns emerged. Impulsivity has previously been found to result in increased susceptibility to phishing emails, specifically that individuals who are less impulsive are more vulnerable to certain types of phishing attacks (Kumaraguru et al., 2007a). Experiment 3 found that the more impulsive (i.e., from BIS-11 scores) an individual was, the more likely they were to rate their task as difficult. These two impulsivity findings make it difficult to interpret how impulsivity plays a role in phishing detection, both potentially suggesting that impulsive individuals are more aware of their limitations and less vulnerable.

However, other research (Parsons et al., 2013) has found that impulsivity negatively impacts phishing detection. Despite these conflicting accounts, it is interesting to note that neither impulsivity nor inhibitory control seem to be related to time spent to classify emails. Thus, more work is necessary to elucidate the true influence of deficient self-regulation on phishing vulnerability.

A personality measure was also related to performance in the cyber task, agreeableness. Agreeableness was originally hypothesized to be linked to the increased utilization of the interventions. However, agreeableness was only related to the perceived threat of phishing emails. Specifically, that individuals who are more agreeable were less likely to perceive phishing emails as threatening. This is an important finding that indicates agreeable individuals may be more vulnerable to phishing emails, potentially due to misplaced trust towards emails. Previous research (Barrick & Mount, 1991) has suggested that agreeable individuals are more trusting. In the context of phishing this could result in a decreased ability to detect fraudulent emails due to over trust. Trust has been linked to phishing susceptibility in previous research. Specifically, that distrust in the senders of phishing emails results in increased classification accuracy (Welk et al., 2015; Wright & Marett, 2010) and trust in the sender results in increased vulnerability (Martin, Lee, & Parmar, 2019).

More specific individual difference variables related to cyber behaviors were also included in the present studies. Individuals who had better cyber hygiene took longer to classify phishing emails in Experiment 3 and were more likely to detect the phishing emails in Experiment 1. This suggests that general safe online behaviors are linked to the ability to detect phishing emails. Although these results are limited in their causal inferences, they do suggest that further training and intervention studies that focus on general safe online behaviors may be able

to improve phishing detection. Additionally, both cyber hygiene and cyber experience were found to be linked to difficulty ratings for legitimate and phishing emails in Experiment 3. Specifically, that individuals who reported better cyber hygiene and more cyber experience found the email task less difficult. Lastly, in Experiment 2, individuals who had more cyber experience found legitimate emails more threatening, possibly demonstrating an increase in awareness of cyber threats. This is consistent with previous research that has found cyber knowledge and experience to be linked with increased resilience to phishing attacks (Harrison, Svetieva, & Vishwanath, 2016). Although the present results are limited in their causal inferences, they do suggest that further training and intervention studies that focus on general safe online behaviors and experience with phishing emails may be able to improve phishing detection.

Vulnerability to Phishing Emails

A consistent theme across all four experiments was the overwhelming poor email classification performance. Although accuracy was higher for legitimate emails in Experiment 1, phishing email detection was near chance performance. Email classification accuracies remained low across the remaining three experiments, and all sensitivities (d') fell below 1.5, indicating that all participants regardless of the experiment struggled to classify emails. Additionally, many participants were liberal in their classifications, classifying more emails as legitimate than phishing. This bias is particularly concerning given how low the sensitivities were. Even more troubling than their classification accuracies were the inappropriate actions participants selected for phishing emails. On roughly 20% of phishing emails participants said the next action they would take was to click a link/open an attachment or reply. These types of actions would result

in an email user compromising their personal information in the real world. Participants also demonstrated poor metacognition for their performance on the cyber task. Participants did rate phishing emails as more threatening that legitimate emails across the experiments, but only rated the phishing emails as mildly threatening. This perceived threat level should be much higher, given that any of these phishing emails could have potentially stolen their personal information (e.g., social security numbers, credit card information) if the participants interacted with them in the real world. Participants were also highly confident and viewed the task as relatively easy despite their poor task performance. This miscalibration of confidence and ability is consistent with previous studies related to metacognition and multitasking (e.g., Ophir, Nass, & Wagner, 2009; Sanbonmatsu, Strayer, Medeiros-Ward, & Watson, 2013) and recent phishing studies (Canfield, Fischhoff & Davis, 2019). Overall, participants appeared to exhibit extremely poor performance across the board with little awareness of their vulnerabilities.

How to Improve Performance

The present studies have established just how bad email users can be at classifying emails under various task factors. However, little work has demonstrated how we can improve a user's resilience to fraudulent email attacks. The first three experiments suggest that longer classification times may result in improved detection of phishing emails. Specifically, in Experiment 1 there was a positive relationship between response times and phishing classification accuracy. Additionally, the best performance in Experiment 2 was seen for the 25 phishing email condition. This group of participants took longer to evaluate phishing emails and this seems to have contributed to their higher overall accuracy. Participants also tended to take longer to classify phishing emails compared to legitimate emails in Experiment 3. Previous

research has also found that response times are linked to performance, with quicker judgments resulting in decreased phishing detection (Jones, Towse, Race & Harrison, 2019). Together these findings suggest that one way to improve phishing detection is to evaluate emails slowly. Email interfaces like GMAIL and Outlook could implement changes where users can't interact (e.g., respond, clicking, opening attachments) with an email until a certain time limit has passed. This might prevent any impulsive and/or motor errors that could occur when engaging with emails.

The main aim of Experiment 4 was to develop novel interventions to increase email users' resilience to fraudulent emails. As predicted, the embedded cheat sheet resulted in the best performance relative to the control group with a roughly 20% boost in phishing detection performance. The embedded cheat sheet group from Experiment 4 also selected the safest actions across all the experiments with roughly 90% accuracy. Additionally, although not statistically different from the other groups, the embedded cheat sheet group appeared to be less biased with their response criterion being close to 0. Although the embedded cheat sheet group did demonstrate several benefits, their overall accuracy and sensitivity was still poorer than what should be considered acceptable. For an intervention to be implemented it needs to reliably increase classifications to near ceiling performance. Interestingly, neither intervention seemed to change response times, perceived threat level, confidence or difficulty. Embedded training has also recently been linked to improved phishing detection in webpages (Xiong, Proctor, Yang, & Li, 2019) and has been previously linked with better performance with emails (Kumaraguru et al., 2007b). Overall, embedded information appears to be a viable avenue for future intervention. However, further research is necessary to develop a more robust email intervention that can improve classification accuracy to near ceiling performance.

Limitations and Future Directions

Although the present studies contributed to the cyber domain's understanding of susceptibility to phishing emails, there are several limitations and areas for future research. One limitation of the present results is the addition of the timer in Experiment 3. In order to exacerbate the effects of time pressure this timer was introduced part way through the experiments. However, this methodological difference does make it difficult to directly compare performance between Experiments 1,2, and 3. It is possible that this additional time pressure manipulation washed out any effects of prevalence in Experiment 3. Prevalence effects may have also been challenging to find because of how poorly the participants performed. In some cases, higher prevalence conditions may have only decreased performance since participants had more opportunities to miss phishing emails due to their low discernibility (i.e., sensitivities) and bias towards saying emails were from legitimate sources.

Additionally, in Experiment 4 even though performance improved in the embedded cheat sheet conditions, phishing detection was nowhere near optimal levels. It is possible that greater performance improvements would be seen with additional interventions such as feedback and information about the dangers of interacting with cyber-attacks. Future research is necessary in order to determine what interventions will result in the most benefits to phishing detection. Lastly, we found very minimal relationships between our individual difference variables and dependent measures. It is possible that these analyses were just underpowered but it is also possible we had a limited sample. For example, the cyber experience data was constricted to low levels of previous cyber experience, making it difficult to find any meaningful relationships. More studies are necessary that have experimental control over these types of variables (i.e.,

recruiting cyber experts and novices) to more fully understand these relationships. Overall, the present studies indicated that task factors such as email load and phishing prevalence can decrease fraud detection, but that embedded phishing tips can improve performance.

APPENDIX A: EXPERIMENT 1 APPROVAL LETTER

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

EXEMPTION DETERMINATION

May 19, 2019

Dear Dawn Samo:

On 5/19/2019, the IRB determined the following submission to be human subjects research that is exempt from regulation:

Type of Review:	Initial Study, Category
Title:	Email Classifications and Personality
Investigator:	Dawn Samo
IRB ID:	STUDY00000512
Funding:	None
Grant ID:	None

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made, and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request so that IRB records will be accurate.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Kamille Chaparro
Designated Reviewer

APPENDIX B: EXPERIMENT 2 APPROVAL LETTER

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

EXEMPTION DETERMINATION

June 21, 2019

Dear Dawn Sarno:

On 6/21/2019, the IRB determined the following submission to be human subjects research that is exempt from regulation:

Type of Review:	Initial Study, Exempt Category
Title:	Email Classifications and Previous Experience
Investigator:	Dawn Sarno
IRB ID:	STUDY00000600
Funding:	None
Grant ID:	None

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made, and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request so that IRB records will be accurate.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Kamille Chaparro
Designated Reviewer

APPENDIX C: EXPERIMENT 3 APPROVAL LETTER

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

EXEMPTION DETERMINATION

August 13, 2019

Dear Dawn Sarno:

On 8/13/2019, the IRB determined the following submission to be human subjects research that is exempt from regulation:

Type of Review:	Initial Study, Exempt Category
Title:	Email Classifications and Actions
Investigator:	Dawn Sarno
IRB ID:	STUDY00000742
Funding:	None
Grant ID:	None

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made, and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request so that IRB records will be accurate.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Kamille Chaparro
Designated Reviewer

APPENDIX D: EXPERIMENT 4 APPROVAL LETTER

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

EXEMPTION DETERMINATION

August 13, 2019

Dear Dawn Sarno:

On 8/13/2019, the IRB determined the following submission to be human subjects research that is exempt from regulation:

Type of Review:	Initial Study, Exempt Category
Title:	Email Classifications and Performance
Investigator:	Dawn Sarno
IRB ID:	STUDY00000743
Funding:	None
Grant ID:	None

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made, and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request so that IRB records will be accurate.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Kamille Chaparro
Designated Reviewer

REFERENCES

Baddeley, A. D., & Colquhoun, W. P. (1969). Signal probability and vigilance: A reappraisal of the 'Signal Rate' effect. *British Journal of Psychology*, *60*(2), 169-178.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, *44*(1), 1-26.

Bergholz, A., De Beer, J., Glahn, S., Moens, M. F., Paab, G., & Strobel, S. (2010). New filtering approaches for phishing email. *Journal of computer security*, *18*(1), 7-35.

Byrne, Z. S., Dvorak, K. J., Peters, J. M., Ray, I., Howe, A., & Sanchez, D. (2016). From the user's perspective: Perceptions of risk relative to benefit associated with using the Internet. *Computers in Human Behavior*, *59*, 456-468.

Cain, A. A., Edwards, M. E., & Still, J. D. (2018). An exploratory study of cyber hygiene behaviors and knowledge. *Journal of information security and applications*, *42*, 36-45.

Canfield, C. I., Fischhoff, B., & Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human factors*, *58*(8), 1158-1172.

Canfield, C. I., Fischhoff, B., & Davis, A. (2019). Better beware: comparing metacognition for phishing and legitimate emails. *Metacognition and Learning*, *14*(3), 343-362.

Chandrasekaran, M., Narayanan, K., & Upadhyaya, S. (2006). Phishing email detection based on structural properties. *NYS cyber security conference* (Vol. 3).

Coutlee, C. G., Politzer, C. S., Hoyle, R. H., & Huettel, S. A. (2014). An Abbreviated Impulsiveness Scale constructed through confirmatory factor analysis of the Barratt Impulsiveness Scale Version 11. *Archives of scientific psychology*, *2*(1), 1.

Downs, J. S., Holbrook, M. B., & Cranor, L. F. (2006). Decision strategies and susceptibility to phishing. *Proceedings of the second symposium on Usable privacy and security* (pp. 79-90). ACM.

Drake, C. E., Oliver, J. J., & Koontz, E. J. (2004). Anatomy of a Phishing Email. CEAS.

Elkind, P. (2015) Inside the hack of the century. *Fortune*. Retrieved from: http://fortune.com/sony-hack-part-1/

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175-191. doi: 10.3758/BF03193146

Ferguson, A. (2005). *Fostering E-mail Security Awareness: The West Point Carronade*. Retrieved from: https://er.educause.edu/articles/2005/1/fostering-email-security-awareness-the-west-point-carronade.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. *The International World Wide Web Conference*.

Frederick, S. (2005) Cognitive reflection and decision making. *Journal of Economic Perspectives, 19,* 4, 25–42.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. Los Altos, CA : Peninsula Pub., 1988.

Grimes, G.A., Hough, M.G., & Signorella, M.L. (2007). Email end users and spam: relations of gender and age group to attitudes and actions. *Computers in Human Behavior*, *23*¸ 318-332.

Hadlington, L. (2017). Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours. *Heliyon*, *3*(7),1-18.

Harrison, B., Svetieva, E., & Vishwanath, A. (2016). Individual processing of phishing emails: How attention and elaboration protect against phishing. *Online Information Review*, *40*(2), 265-281.

Jakobsson, M. (2007). The human factor in phishing. *Privacy & Security of Consumer Information*, *7*(1), 1-19.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The Big Five Inventory—Versions 4a and 54. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Jones, H. S., Towse, J. N., Race, N., & Harrison, T. (2019). Email fraud: The search for psychological predictors of susceptibility. *PloS one*, *14*(1), e0209684.

Kircanski, K., Notthoff, N., DeLiema, M., Samanez-Larkin, G. R., Shadel, D., Mottola, G., ... & Gotlib, I. H. (2018). Emotional arousal may increase susceptibility to fraud in older and younger adults. *Psychology and aging*, *33*(2), 325.

Kumaraguru, P., Rhee, Y., Sheng, S., Hasan, S., Acquisti, A., Cranor, L.F., and Hong, J. (2007a) Getting users to pay attention to anti-phishing: Evaluation of retention and transfer. In the *Proceedings of the e-Crime Researchers Summit, Anti-Phishing Working Group* (pp.70-81). New York, NY:ACM.

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007b). Protecting people from phishing: the design and evaluation of an embedded training email system. In the *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 905-914). New York, NY: ACM.

Levine, T. R., & McCornack, S. A. (1991). The dark side of trust: Conceptualizing and measuring types of communicative suspicion. *Communication Quarterly*, *39*(4), 325-340.

Logan, G. D., Schachar, R. J., & Tannock, R. (1997). Impulsivity and inhibitory control. *Psychological science*, *8*(1), 60-64.

Macworth, N.H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6-21.

Martin, S. R., Lee, J. J., & Parmar, B. L. (2019). Social distance, trust and getting "hooked": A phishing expedition. *Organizational Behavior and Human Decision Processes*.

Mayhorn, C. B., & Nyeste, P. G. (2012). Training users to counteract phishing. *Work*, *41*, 3549-3552.

McBride, M., Carter, L., & Warkentin, M. (2012). Exploring the role of individual employee characteristics and personality on employee compliance with cybersecurity policies. *RTI International-Institute for Homeland Security Solutions*.

Ophir, E,, Nass, C., Wagner, A.D. (2009). Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences*. *106*(37).

Parasuraman, R., Hancock, P. A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centered collision-warning systems. *Ergonomics*, *40*(3), 390-399.

Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., & Jerram, C. (2013). Phishing for the truth: A scenario-based experiment of users' behavioural response to emails. *IFIP International Information Security Conference* (pp. 366-378). Springer, Berlin, Heidelberg.

Patel, P., Sarno, D., Lewis, J. E., Neider, M. B., & Bohil, C. J. (In press). Perceptual representations of spam and phishing emails. *Applied Cognitive Psychology*.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of clinical psychology*, *51*(6), 768-774.

Plummer, M. (2019). *How to spend way less time on email every day*. Retrieved from: https://hbr.org/2019/01/how-to-spend-way-less-time-on-email-every-day

Sanbonmatsu. D.M., Strayer, D.L., Medeiros-Ward, N., Watson, J.M. (2013) Who multi-tasks and why? Multitasking ability, perceived multi-tasking ability, impulsivity, and sensation seeking. *PLoS ONE*. *8*(1).

Sarno, D. M., Lewis, J. E., Bohil, C. J., Shoss, M. K., & Neider, M. B. (2017). Who are phishers luring?: A demographic analysis of those susceptible to fake emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. *61* (1), 1735-1739.

Sarno, D. M., Lewis, J. E., Bohil, C. J., & Neider, M. B. (in press). Phishing through the Generations: Cybersecurity Threats across the Lifespan. *Human Factors.*

Sawyer, B. D., Finomore, V. S., Funke, G. J., Mancuso, V. F., Funke, M. E., Matthews, G., & Warm, J. S. (2014). Cyber vigilance: effects of signal probability and event rate. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58* (1), 1771-1775.

Sawyer, B. D., Finomore, V. S., Funke, G. J., Mancuso, V. F., Miller, B., Warm, J., & Hancock, P. A. (2015). Evaluating cybersecurity vulnerabilities with the email testbed: Effects of training. *Proceedings 19th Triennial Congress of the IEA*, *9*, 14.

Sawyer, B. D., & Hancock, P. A. (2018). Hacking the Human: The Prevalence Paradox in Cybersecurity. *Human factors*, *60*(5), 597-609.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science*, *3*(4), 207-218.

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L., & Downs, J. (2011). Who falls for phish?: A demographic analysis of phishing susceptibility and effectiveness of interventions. *Conference on Human Factors in Computing Systems Proceedings*, 373-382.

Shropshire, J., Warkentin, M., Johnston, A., & Schmidt, M. (2006). Personality and IT security: An application of the five-factor model. *AMCIS 2006 Proceedings*, 415.

Shropshire, J., Warkentin, M., & Sharma, S. (2015). Personality, attitudes, and intentions: Predicting initial adoption of information security behavior. *Computers & Security*, *49*, 177-191.

Silva, A., Emmanuel, G., McClain, J. T., Matzen, L., & Forsythe, C. (2015). Measuring expert and novice performance within computer security incident response teams. *Foundations of Augmented Cognition, 9183*, 144-152.

Singh, K., Aggarwal, P., Rajivan, P., & Gonzalez, C. (2019). Training to detect phishing Emails: effects of the frequency of experienced phishing emails. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63* (1), 453-457.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, *31*(1), 137-149.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

The Council of Economic Advisers (2018). *The Cost of Malicious Cyber Activity*. Retrieved from: https://www.whitehouse.gov/wp-content/uploads/2018/03/The-Cost-of-Malicious-Cyber-Activity-to-the-U.S.-Economy.pdf.

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.

Vishwanath, A., Harrison, B., & Ng, Y. J. (2016). Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research*, 1-21.

Wang, J., Herath, T., Chen, R., Vishwanath, A., & Rao, R. (2012). Phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE Transactions on Profession Communication*, *55* (4), 345-362.

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?: Assessing individual differences in a phishing detection task. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL)*, *5*(4), 1-17.

Williams, S., Sarno, D., Lewis, J., Shoss, M., Neider, M., & Bohil, C. (2019). The Psychological interaction of spam email features. *Ergonomics*, *62*(8), 983-994.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Cognitive psychology: rare items often missed in visual searches. *Nature*, *435*(7041), 439-443.

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.

Xiong, A., Proctor, R. W., Yang, W., & Li, N. (2019). Embedding Training Within Warnings Improves Skills of Identifying Phishing Webpages. *Human factors*, *61*(4), 577-595.