A Phish Scale: Rating Human Phishing Message Detection Difficulty

Michelle P. Steves
National Institute of Standards and Technology
michelle.steves@nist.gov

Kristen K. Greene
National Institute of Standards and Technology
kristen.greene@nist.gov

Mary F. Theofanos
National Institute of Standards and Technology
mary.theofanos@nist.gov

***Abstract*—As organizations continue to invest in phishing awareness training programs, many Chief Information Security Officers (CISOs) are concerned when their training exercise click rates are high or variable, as they must justify training budgets to those who question the efficacy of training when click rates are not declining. We argue that click rates should be expected to vary based on the difficulty of the phishing email for a target audience. Past research has shown that when the premise of a phishing email aligns with a user's work context, it is much more challenging for users to detect a phish. Given this, we propose a Phish Scale, so CISOs and phishing training implementers can easily rate the difficulty of their phishing exercises and help explain associated click rates. We based our scale on past research in phishing cues and user context, and applied it to previously published data and new data from organization-wide phishing exercises targeting approximately 5 000 employees. The Phish Scale performed well with the current phishing dataset, but future work is needed to validate it with a larger variety of phishing emails. The Phish Scale shows great promise as a tool to help frame data sharing on phishing exercise click rates across sectors.**

Keywords—phishing cues, embedded phishing awareness training, operational data, network security, phishing defenses, security defenses

I. Introduction

According to Cybersecurity Ventures' 2017 Official Annual Cybercrime Report, it is estimated that cybercrime damages will cost the world $6 trillion annually by 2021 [5]. These cost projections are supported by historical cybercrime figures and recent year-over-year growth. Furthermore, there has been a notable increase in hacking activities sponsored by hostile nation states, as well as activities from organized crime syndicates. Finally, the cyber attack surface continues to grow, in large part due to an explosion of Internet of Things (IoT) devices. Humans are another particularly important component of the overall attack surface, as social engineering continues to be successful. In recognition of the importance of human behavior in cybersecurity, organizations are more widely investing in cybersecurity awareness programs for their computer users, and often on phishing training in particular. Embedded phishing awareness training is popular—and in some cases, mandated—in a wide variety of sectors, such as financial services, government, healthcare, and academia. In this type of training, simulated phishing emails are sent that mimic real-world threats, in order to raise employee phishing awareness.

Workshop on Usable Security (USEC) 2019
24 February 2019, San Diego, CA, USA
ISBN 1-891562-57-6
https://dx.doi.org/10.14722/usec.2019.23028
www.ndss-symposium.org

Not surprisingly, many Chief Information Security Officers (CISOs) are concerned when their training exercise click rates are high. This is especially true for more mature or long-running awareness programs, as CISOs often expect lower and lower click rates to show the effectiveness of training. Further, the Return on Investment (ROI) for such training may be questioned if click rates are high or even variable. However, low click rates do not necessarily indicate training effectiveness and may instead mean the phishing emails used were: 1) too easy, 2) not contextually relevant for most staff, or 3) the phish was repeated or very similar to previous exercises. In fact, low click rates and training programs in general, can generate a false sense of security or complacency if considered in isolation. Phishing awareness training program click rates must be part of a more comprehensive, metrics-informed approach to effectively understand and combat phishing threats [15].

Past work [14] has shown that click rates will vary based on the contextual relevance of the phish, with highly contextually relevant phish resulting in extreme spikes in click rates—despite years of phishing awareness training. Furthermore, attackers continue to refine and vary phishing attack premises. Although "traditional" phishing emails are still quite successful, attackers are becoming more sophisticated and creative all the time. Additionally, there is a treasure trove of readily available information online that attackers can use to better tailor phish and capitalize on contextual relevance. While some information is willingly and openly shared by users on social media, much other information has been exposed through large-scale data breaches, such as the recent Facebook hack [29].

While repetition is important for training phishing recognition and for conditioning reporting behavior, simple repetition of the same or very similar phishing emails does not represent the full spectrum of phishing threats observed in the real world. It is important to vary phishing exercises appropriately and challenge staff with contextually relevant phish of varying difficulty to provide training on new scams—for which variable click rates should be expected. This should not be viewed as a negative effect but rather a positive outcome, as it means organizations are truly training their staff with phish that represent current real-world threats. But how exactly does one measure the difficulty of a given phishing email? While we can certainly measure click rates post-hoc and infer detection difficulty somewhat from those numbers, we would prefer an a priori method of difficulty determination. In discussions with CISOs at [17], [18], and others, we found that a method to determine phishing message difficulty would indeed be highly beneficial for those responsible for phishing training implementation. To meet this need, we propose a Phish Scale, an easy way for CISOs and training implementers to characterize the difficulty of their phishing exercises and provide context for the associated metrics. This context is a missing element that training implementers need to improve the training benefit of their exercises and subsequent ROI.

In this paper, we describe our exploratory effort to construct a preliminary conceptual version of the Phish Scale and its components. Further, we use the Phish Scale to determine the difficulty rating for each of seven real-world phishing training exercises that are described in detail. Then we observe if the Phish Scale difficulty rating for each exercise aligns with the exercise's actual click rate. Finally, we discuss our observations, limitations of the effort to-date, as well as future work including refinement of its components and validation of the overall scale through a wider variety of phishing emails.

II. Background

Technological and human-centered approaches are used in conjunction to combat email phishing. Technologically-focused approaches include mechanisms like filtering, firewalls, and blacklists, whereas human-centered approaches tend to focus on cybersecurity awareness training, and often on phishing specifically. Due in part to advances in Machine Learning (ML) and Artificial Intelligence (AI), email filters in particular, are becoming ever more effective at blocking generic spam. This has meant that users now see fewer emails of this nature in their inboxes. Recent work has posited the existence of the "Prevalence Paradox" [34], suggesting that users may therefore be more vulnerable when such emails do get through, due to their reduced experience with potentially malicious emails. Yet other work [25] has shown that people often expand their concept of a given stimulus in response to a decrease in the prevalence of said stimulus, for example, seeing neutral faces as threatening when threatening faces became rare. Although the series of experiments by Levari et al. did not address phishing specifically, given the set of topics they investigated, it would certainly be plausible to expect their findings to hold in the phishing domain. We hope additional research on the effects of prevalence on phishing detection—for both humans and AI—will reconcile different findings on prevalence.

In addition to prevalence, there are numerous other factors that complicate human detection of phishing emails. There are several existing theories and models of phishing susceptibility that are highly relevant for the development of a Phish Scale. These theories and models directly address the types of email cues, tactics, and individual user characteristics that together help—at least partially—explain the relative ease or difficulty of human phishing detection.

Protection Motivation Theory, or PMT [33] addressed user perceptions of threat and corresponding perceived threat management ability. PMT has largely been applied to security behavior in general, although Wang et al. [41] did apply PMT specifically to phishing threat perception. Much more recently than PMT, which was originally proposed in 1975, an Integrated Information Processing Model of Phishing Susceptibility, or IIPM, was proposed [38]. The IIPM proposed that users' limited attentional resources for information processing are essentially hijacked when certain techniques like urgency are used to influence behavior, meaning that users rely on heuristic information processing (System I, [22]), rather than engaging in deeper, more systematic processing (System II, [22]). When this type of surface level information processing style is used it makes users more likely to overlook or ignore cues that might otherwise tip a user off as to the legitimacy of the email, such as an incorrect sender address. In 2016, Vishwanath et al. proposed the Suspicion, Cognition, and Automaticity Model, or SCAM, which posited that individual user characteristics cause variability in the use of heuristic processes for email evaluation [39].

Recent work by Williams, Hinds, and Joinson [42] considered these three models (PMT, IIPM, and SCAM) within the work context of an international organization with sites in the UK, finding that the presence of authority cues increased the likelihood that users would click a suspicious email link [42]. In addition to the types of models or theories such as PMT, IIPM, SCAM, there is a large wealth of prior work investigating or describing the impact of particular email cues, such as inclusion of authority and urgency cues. Research on phishing cues is particularly relevant for development of a Phish Scale, as email users rely on cues to determine if a particular email message is a phish.

Indeed, anti-phishing advice and training stress the characteristics of phishing messages that email users should look for; these are often called cues, indicators and hooks. The list of cues is long and varied, such as those contained in [26], [30]. Because there is no set pattern of which cues may be contained in any particular message, the task for users when determining if a message is a phish is harder than if the list were very short. Making the task even more difficult, prior work shows the alignment of the phish's premise and user context affects which cues the user finds to be salient. Further, the same cue can be compelling for some users but suspicion generating for others—depending on the user's context [14], [15].

In the Greene et al. [14] study, phishing exercise data were collected over 4.5 years in an ecologically valid workplace setting, with corresponding survey data for the final year. The study found that user context was extremely important in phishing susceptibility; the authors proposed that it was the lens through which users viewed and interpreted email cues. When a user's work context was misaligned with the premise of the phishing email, they were more likely to attend to suspicious cues. For example, they have no invoicing responsibilities at work and the phishing email was purportedly an unpaid invoice. In contrast, when a user's work context was well aligned with the phishing email premise, they were more likely to attend to compelling cues, and completely ignore or largely discount suspicious cues. In this case, if the user is directly responsible for paying invoices at work and the phishing email was purportedly an unpaid invoice.

Greene et al. [14] emphasized the importance of phishing research in the workplace setting, as much prior phishing work was conducted in laboratories with artifical user contexts or university settings that can be quite different than the workplace. Williams et al. [42] also addressed this need for workplace data in their research. One of the few other studies situated in the workplace was conducted by Caputo et al. [4], but due to limitations was only able to suggest the possible importance of user context. We further contribute to the growing corpus of workplace-based phishing research, by applying our Phish Scale to three previously published workplace-based phishing exercises in [14], as well as, reporting on and applying the Phish Scale to four new workplace-situated phishing exercises. Two of our new exercises have much larger sample sizes, with *n*'s of ~5000 for each exercise, compared to those previously reported in [14], with *n*'s of ~70 for each exercise.

III. Method

To assist those tasked with implementing phishing awareness training programs, it is important to consider the relative detection difficulty of training messages. Phishing messages, whether those intended for training or actual threats, can be more or less difficult *for a given work group* to detect as a phishing attempt. Understanding the detection difficulty helps phishing awareness training implementors in two primary ways: 1) by providing context regarding training message click and reporting rates for a target audience, and 2) by providing a way to characterize actual phishing threats so the training implementor can reduce the organization's security risk by tailoring training to the types of threats their organization is facing. To this end, we are attempting to develop a Phish Scale—to help practitioners rate training messages both to contextualize click rates for embedded phishing awareness training as well as tailor training efforts. We anticipate it will provide CISOs with another metric to help gauge the progress of their awareness programs over time and address risk. The scale is intended to categorize the detection difficulty of a phishing message with respect to a target audience.

In the remainder of this section, we describe the Phish Scale and the operationalization of its components into a single framework. In the next section, we present data from seven workplace-situated phishing awareness training exercises to illustrate how to derive a phish difficulty rating using the Phish Scale. The data were gathered with appropriate human subjects approval at NIST.

A. The Phish Scale

To develop our Phish Scale, we began by considering the primary elements that CISOs and/or training implementors use when selecting and customizing phishing training exercises. These elements are scenario premise and message content. The scenario premise may pertain to a relatively new threat or an older threat that remains effective for a particular target audience. The message content is often customizable by the trainer and contains the cues that trainees might use to detect the training phish. For this exploratory effort, we root the Phish Scale in these two primary elements: the *cues contained in the message* and the *premise alignment for the target audience*.

Other factors such as personality, phishing tactics knowledge, concern for security, concern for consequences, and the like certainly affect click rates, and ultimately we intend to consider incorporating additional factors such as these; we return to this topic in the future work section. However, for now, this effort starts with message cues and premise alignment as these elements undoubtedly play crucial roles in phishing detection by humans and, importantly, they can be categorized by training implementors for a given target audience. For this initial effort at characterizing detection difficulty, the Phish Scale components are:

1) A rating system for observable characteristics of the phishing email itself, such as the number of cues, nature of the cues, repetition of cues, and so on.

2) A rating system for alignment of the phishing email premise with respect to a target audience.

Table I presents our exploratory, conceptual framework illustrating how detection difficulty rating is arrived at once the categories for number of cues and premise alignment are determined. In an attempt to keep the categorization relatively simple for training implementors, we used three categories for each component and assigned labels representing relative ranges for each. Next we discuss the operationalization of each component. In the following section, we walk through marrying the real-world phishing training exercise data with these conceptual categorizations and discuss our observations.

TABLE I: The Phish Scale

Number of Cues	**Premise Alignment**	**Detection Difficulty**
Few (more difficult)	High	Very difficult
	Medium	Very difficult
	Low	Moderately difficult
Some	High	Very difficult
	Medium	Moderately difficult
	Low	Moderately to Least difficult
Many (less difficult)	High	Moderately difficult
	Medium	Moderately difficult
	Low	Least difficult

In the conceptual framework we acknowledge the stronger influence of premise alignment component over cues, as reported in [14]. This is reflected in the detection difficulty rating tending to be at the *Very difficult* or *Moderately difficulty* rating when the premise alignment is categorized as *High* or *Medium*. Additionally, there are more *Very difficult* detection difficulty rating assignments than *Least difficult* rating assignments in the entire conceptual Phish Scale framework. The detection difficulty rating for the combination of *Some* cues and *Low* premise alignment was given a range from *Moderately to Least difficult* rating, further reflecting our belief that even a *Low* premise alignment can have a disproportionate effect on increasing detection difficulty. While we expect all of the ratings to be informed with empricial data, this is especially true for this particular combination (low premise alignment and some cues). Finally, we purposefully did not label a category as *Easy to detect* or similar, as we expect that the premise of any phishing message will typically align for *at least a few users* and for them, detection is often not easy.

Ultimately, we anticipate that each detection difficulty rating will equate to a range of click rates. For example, the phishing training messages that have a corresponding detection difficulty rating of *least difficult* may be expected to have a click rate of less than 10 %. We return to this topic in the discussion after we have examined the empirical data presented in the next section.

B. Phishing message cues

To incorporate the effect of phishing message cues in the scale, we decided to use the count of instances of those characteristics that are present in the message being rated. Our reasoning is that the fewer phishing cues present in a message, the more difficult it is to detect. Conversely, the more cues present, the more opportunities for a user to notice a tip-off that generates suspicion. We realize the effect of any single cue or hook can differ from instance to instance and person to person. Indeed, we return to this topic in the discussion. Currently, there are three categories in the framework to describe the quantity of these characteristics: *Few* (fewer opportunities to detect), *Some*, and *Many* (more opportunities to detect).

Before we can count cues, we needed to determine which phishing characteristics—the list of cues, indicators and hooks—are appropriate for inclusion in the framework. From [14] we see that a particular phishing characteristic may either be suspicion-generating (a tip-off) or compelling (a hook), depending on the user's context. In keeping with prior literature, we use the term "cue." However, we mean it in the broader sense of a phishing message characteristic. We require that each cue included in the framework be able to be tied to an objectively observable characteristic in a message.

From the literature we considered the compendiums of phishing cues in [26] and [30]. We used the cues given in [26] as a starting point. Additionally, we modified the categories in an attempt to order the cues from those that are often suspicion-generating, such as errors, to those that are typically compelling, such as common tactics, these tactics being commonly used because they continue to be compelling. This is a rough ordering of categories at best, but we felt it is better suited to counting cues than those given in [26] and [30]. The categories are: *Error*—relating to spelling and grammar errors and inconsistencies contained in the message; *Technical indicator*—pertaining to email addresses, hyperlinks and attachments; *Visual presentation indicator*—relating to branding, logos, design and formatting; *Language and content*—such as a generic greeting and lack of signer details, use of time pressure and threatening language; and, *Common tactic*—use of humanitarian appeals, too good to be true offers, time-limited offers, poses as a friend, colleague, or authority figure, and so on. We wove in additional phishing characteristics from [14], [30], and others.

Table II provides the list of cues we identified that are objectively present in phishing messages. The table in the Appendix A contains the same list of cues, but is expanded with a brief description of each, associated references, and the criteria we used when deciding if a particular cue was observably present in an individual message. To determine the cues count, use the criteria given in Appendix A for each cue, count how many instances for each and sum for a total.

For this initial effort, we recognize this list is not exhaustive and will be expanded. Additionally, we anticipate some form of weighting will be useful to reflect cue saliency. Given the variability in cue saliency for individuals within a target population, this is a non-trivial exercise. These are refinements we expect will come with additional development of the scale.

For the purpose of the Phish Scale, we did not include phishing message cues related to mismatches with the user's world, such as an individual's particular work responsibilities or an individual's expectations, for example expecting an important phone call. Work responsibilities and general workplace expectations for the target audience are folded into the premise alignment component of the Phish Scale.

TABLE II: PHISHING MESSAGE CUES

Cue Type	Cue Name
Error	Spelling and grammar irregularities
	Inconsistency
Technical indicator	Attachment type
	Sender display name and email address
	URL hyperlinking
	Domain spoofing
Visual presentation indicator	No/minimal branding and logos
	Logo imitation or out-of-date branding/logos
	Unprofessional looking design or formatting
	Security indicators and icons
Language and content	Legal language/copyright info/disclaimers
	Distracting detail
	Requests for sensitive information
	Sense of urgency
	Threatening language
	Generic greeting
	Lack of signer details
Common tactic	Humanitarian appeals
	Too good to be true offers
	You're special
	Limited time offer
	Mimics a work or business process
	Poses as friend, colleague, supervisor, authority figure

C. *Phishing premise alignment*

Incorporating premise alignment is a process of characterizing the pertinence of the email message premise for the target audience. It attempts to capture alignment with the following: work responsibilities and business practice plausibility for the target audience. For example, the organization's current business practices and staff expectations reflecting the organization's workplace culture. The premise alignment is expected to be determined by the training implementor—someone with knowledge of the target audience's work responsibilities and expectations as a group.

We use three categories to characterize the alignment: *High*, *Medium*, and *Low*. To determine premise alignment, the training implementer must understand and categorize the premise relevancy for portions of the target audience using the guidelines below.

1) *High alignment*

For high premise alignment, there should be a significant portion of the target audience for which the premise matches with work responsibilities, is highly plausible, and/or aligns strongly with an audience-relevant event. For example, if the recipient population is the finance department and the phishing message has a premise of a late/missed payment, the overall alignment is high.

2) *Medium alignment*

Medium alignment is achieved with either case: a) when the premise has plausible but weak context alignment with a large portion of the target audience or b) when the premise has moderate context alignment with a small portion of the target audience. For example, if the recipient population mostly works in one physical location and the phishing message has a moderately pertinent premise for the few members of the recipient population who work in another physical location.

3) *Low alignment*

There is low alignment when the premise pertains to a topic that is not relevant or plausible to the target audience. For example, if the recipient population is the finance department and the phishing message premise pertains to a Call for Papers on biotech research or a similarly unrelated topic, the overall alignment is low.

IV. Application of the Phish Scale

In this section we present data from seven phishing training exercises and use the Phish Scale to determine the detection difficulty rating for each exercise. We followed the appropriate human subjects approval process for our institution. All the awareness training exercises were situated in a workplace environment. First, we provide a description of each exercise, its premise alignment rationale, and a brief description of the target audience size and available information with respect to the premise alignment. Then we gather the data in Table III, including the cue counts provided in Appendix B, and show the detection difficulty rating for each exercise side-by-side with the actual click rate.

The first three phishing exercises (new voicemail, unpaid invoice, and order confirmation) were initially reported in Greene et al. [14]. Here we expand their descriptions to count the cues and categorize the premise alignment. The subsequent four phishing exercises (Gmail, weblogs, Valentine, and security token) represent new data.

Note that although we assigned premise alignment category ratings to these exercises—rather than the training implementers—we did so with pertinent input from the training implementers.

1) *Phishing exercise descriptions*

a) *New voicemail*

Message description: The new voicemail phish appeared to be from a fictitious CorpVM (corpvm@webaccess-alert.com). It appeared to be a system-generated email, with the subject line reading, "You have a new voicemail." There was a large black and green banner at the top of the email with the text, "CorpVM" in white. There were no logos present in the email, however, there was a small black footer with "© 2015 CorpVM Inc." in white. The body of the email began, "You have a new voicemail!" centered in bold text, followed by, "From: Unknown Caller, Received: 03/06/2016, Length: 00:52." Below that text was a personalized [Firstname Lastname] line, followed by, "You are receiving this message because we were unable to deliver it voice message did not go through because the voicemail was unavailable at that moment. To listen to this message, please click here. You must have speakers enabled to listen to the message. * The reference number for this message is qvfl_cjl09-9107319601-2125579909-62. The length of transmission was approximately 52 seconds. The receiving machine's ID: YJH35-TW410-F37JZL. Thank you." Finally, the email closed with smaller text in italic that read, "This is a system-generated message from a send-only address. Please do not reply to this email."

Premise alignment: The alignment is categorized as Medium—the premise was plausible; around the same time as the exercise, a new business process for voicemail notification, *not* delivery, was being rolled out, although without much fanfare. Even though the premise was plausible, it had no or weak context alignment for most, although not all, of the target audience based on survey feedback reported in [14].

Target audience: One Operational Unit (OU) within NIST, handling financial matters (ordering and invoice reconciliation), administrative program support, and technical program support. $n = 69$

b) *Unpaid invoice*

Message description: The unpaid invoice phish appeared to be from a fictitious employee of the same institution as the email recipients, a fellow Federal employee named Jill Preston (jill.preston@nist.gov). The subject line was, "Unpaid invoice #4806." The greeting was personalized with "Dear [Firstname Lastname]." The email body said, "Please see the attached invoice (.doc) and remit payment according to the terms listed at the bottom of the invoice. Let us know if you have any questions. We greatly appreciate your prompt attention to this matter!" The email simply closed with the name "Jill Preston." There was no other contact information included below the name. Of note, there was a file extension mismatch between the way the attachment was referred to in the body of the email (as a *.doc*) and the way the attachment itself was labeled, it appeared to be a *.zip*, with the filename, "invoice_S-37644806.zip". The unpaid invoice phish mimicked the Locky ransomware [32], a real-world threat current at that time.

Premise alignment: The alignment is categorized as High—the premise aligned extremely highly for roughly a third of the target audience and aligned somewhat for the remainder of the department. Additionally, the whole of the targeted OU was on alert for any unpaid invoices following a recent event surrounding a legitimate unpaid invoice.

Target audience: One OU within NIST, handling financial matters (ordering and invoice reconciliation), administrative program support, and technical program support. $n = 73$

c) *Order confirmation*

Message description: The order confirmation phish appeared to be from, "Order Confirmation" (auto-confirm@discontcomputers.com). Note the misspelling of "discount" in the email address. The subject line was personalized and said, "[Firstname Lastname]Your order has been processed," with a space missing between the user's last name and the word "Your." At the top of the email was an image of several holiday packages, with the words, "Order Confirmation" in bold immediately below the holiday package image. There was no personalization in the body of the email, nor was there a greeting of any type. The email body text said, “Thank you for ordering with us. Your order has been processed. We’ll send a confirmation e-mail when your item ships.” This was followed by the words, “Order Details” in orange with, “Order: #SGH-2548883-2619437” (the order number was in blue text). The next section of the email said, “Estimated Delivery Date: 12/02/2016” (the date was in green text), “Subtotal: $59.97,” “Estimated Tax: $4.05,” and “Order Total: $64.02” in bold. There was a large yellow button labeled with the text, “Manage order.” The button was followed by the text, “Thank you for your order. We hope you return soon for more amazing deals.” Near the bottom of the email was an image of a holiday snow globe and the text, “Need it in time for the holidays? Order before December 23 for free over-night shipping.” (“December 23” was in blue). Much smaller gray text below that said, “Unless otherwise stated, items sold are subject to sales tax in in accordance with local laws. For more information, please view tax information.” (“tax information” was in blue). Note the repeated word “in in,” a subtle mistake that is very difficult for users to notice, especially given the small gray font. Finally, at the very bottom of the email appeared three additional links, all in blue on a single line: “Return Policy | Privacy | Account.”

Premise alignment: The alignment is categorized as Medium —the premise aligned for those who had purchasing authority in the OU and for those who had recently placed an order, a small subset of the whole OU. However, the training exercise took place in December, when many people make on-line purchases for the holidays.

Target audience: One OU within NIST, handling financial matters (ordering and invoice reconciliation), administrative program support, and technical program support. *n* = 66

d) Gmail

Message description: The Gmail phish was a particularly clever spear phish. It targeted mid-level management using a spoofed upper management Gmail address, a tactic based on a real-world phish previously observed at NIST. It appeared to come from the personal Gmail account of NIST’s director (firstname.lastname1@gmail.com) and went to a list of laboratory managers. The subject line was, “Safety Awareness,” which is important given that NIST has a very strong emphasis on fostering a culture of safety. The email was personalized with the recipient’s first name. The body said, “Please make sure your groups are aware of this new requirement:” with a link following this text. The email was signed simply with the first name of the organization’s director.

Premise alignment: The alignment is categorized as High—the premise alignment is very strong given the larger organization’s substantial emphasis on workplace safety and that the message appeared to come from NIST’s director—a notable authority figure in this context. Alignment is further strengthened by the target department’s responsibility for the larger organization’s occupational health and safety.

Target audience: One OU within NIST, handling financial matters (ordering and invoice reconciliation), administrative program support, and technical program support. *n* = 64

e) Weblogs

Message description: The weblogs phish was another spear phish. It appeared to come from a system administrator with the email address, notice@nist.gov. The subject line was, “Unauthorized Web Site Access.” There was no personalization. The body said, “*This is an automated email* Our regulators require we monitor and restrict certain website access due to content. The filter system flagged your computer as one that has viewed or logged into websites hosting restricted content. The system is not fool-proof and may incorrectly flag restricted content. The IT department does not investigate every web filter report, but disciplinary action may be taken.” In bold, it said, “Log into the filter system with your network credentials immediately and review your logs to see which websites triggered this alert.” This was followed by a link that was labeled, “Web Security Logs.” There was no contact information given, and the email closed with, “Do not reply to this email. This email was automatically generated to inform you of a violation of our security and content policies.”

Premise alignment: The alignment is categorized as High—the premise aligns with the fact that accessing inappropriate content is indeed a violation of the organization’s Rules of Conduct policy and can be grounds for dismissal for anyone at the organization. The premise capitalizes on the fact that many organizations, including NIST, scan log data routinely. The threat component coupled with the severity of the consequences increases the alignment. Of note, all new employees receive in-person training regarding the organization’s Rules of Conduct and IT policies, where the disciplinary actions associated with inappropriate web content viewing are highly stressed.

Target audience: One OU within NIST, handling financial matters (ordering and invoice reconciliation), administrative program support, and technical program support. *n* = 73

f) Valentine

Message description: The Valentine phish appeared to come from “eCard Delivery” with the email address, “do_not_reply@ecardalert.com.” The subject line said, “Happy Valentine’s Day! See who sent you an e-card…” There were three large red heart images at the top of the email. There was no personalization. The body of the email said, “A secret admirer wished you a Happy Valentine’s Day! Some of you may have heard about our employee greeting cards that can be used to acknowledge fellow employees. Click on the link below to view yours.” This was followed by a large link that said, “Your Card is Waiting,” and additional text that said, “If you are having trouble viewing the e-card please click here.” “Would you like to send an e-card? Visit our site. *Making someone’s day, one e-card at a time…”* The email closed with, “This email may contain confidential and privileged information for the sole use of the intended recipient. If you are not the intended recipient, please contact the sender and delete all copies. Any review or distribution by others is strictly prohibited. Thank you.” The Valentine phish was sent January 22, 2018, prior to Valentine’s Day.

Premise alignment: The alignment is categorized as Low—the premise does not align with a business process but is more personal in nature. However, the message contains a sentence about using the service to acknowledge a fellow employee. So, while the premise does not align with a business practice, it does play on the reader's curiosity, aligning with the upcoming occasion of Valentine's Day, of which most people are aware.

Target audience: All staff at NIST with an email address were targeted, from the human resources department, to finance, to bench scientists, to administrative support and all levels of management. $n = 4977$

g) Security token

Message description: The security token phish appeared to be from "Alerts" with the email address, "alerts@verifytoken.com." The subject line was, "Verify Your Security Token Was Not Compromised." The email was personalized using the format, "Lastname, Firstname, Middle Initial (Fed)." The body said, "Recently we have been made aware of a security breach in our security token product. Some of the tokens have been compromised and may need to be replaced. In order to find out if you [sic] token has been compromised, Validate Your Security Token Here." (Note the "you token" instead of "your token" here.) The email was signed "Rivest Shamir Adleman, Director of Identity and Access Management." The email closed with smaller text that said, "This email may contain confidential and privileged information for the sole use of the intended recipient. Any review or distribution by others is strictly prohibited. If you are not the intended recipient, please contact the sender and delete all copies. Thank you." If someone clicked the "Validate Your Security Token Here" link, they were taken to a data entry webpage, with the URL "secure.verifytoken.com." The top of the webpage said, Token Security with a red background, followed by "Attention!!! Recently the safety of some security tokens has been compromised. Enter your username and six-digit number that is generated every 60 seconds by your security token and we will know if you will need a new token. Should you need a new token, you will be given contact information to request a new token, which will be shipped to you overnight." This was followed by the text, "Account Login," with fields labeled, "User ID" and "Password or Passcode." There was a blue button labeled, "Login" and an "I'm not a robot" checkbox. At the bottom of the webpage was the text, "A passcode contains a PIN and a number from a security token."

Premise alignment: The alignment is categorized as Medium —the premise does not align at all for those personnel who do not have a security token (roughly 43 % of all staff at the organization). Further, the premise does not align for those staff who expect any token checking and replacement would be conducted via the organization rather than a third party—likely a significant portion of the remaining 57 % as the organization has a very strong posture regarding IT security.

Target audience: All staff at NIST with an email address were targeted, from the human resources department, to finance, to bench scientists, to administrative support and all levels of management. $n = 5024$

2) Determining difficulty ratings

a) Applying the Phish Scale

As described previously, the difficulty rating for an individual phishing message is determined first by categorizing the number of objectively observed cues and the premise alignment. Then use the conceptual framework in Table I to select the difficulty rating associated with the categorized number of cues and premise alignment. In Table III the Phish Scale ratings are shown for each of the seven previously described phishing exercises, including the number of cues for each email (detail provided in Appendix B), the premise alignment (from the exercise description), the difficulty rating (from the conceptual framework in Table I), and the actual click rates for each exercise.

The table in Appendix B contains the counts for each cue and a total count for each exercise. When counting cues in a given email message during analysis, it is important to note that these cue counts are based on our extremely careful scrutiny of the email messages; most email users are not going to notice or attend to all the available cues.

Note that in order to calculate the difficulty rating, the number of cues must be further categorized into *Few*, *Some*, or *Many*, in order of decreasing difficulty. Although some cues are more salient than others, we anticipate this is a reasonable first approximation. In this initial version of the Phish Scale, we propose the associated ranges as follows: the category labeled *Few* is represented by 1 to 8 cues, the category labeled *Some* by 9 to 14 cues, and the category labeled *Many* by 15 or more cues. These ranges are based on our existing dataset; at this stage of scale development, the click rates inform the categorization of the cue counts. We fully expect the cue count ranges may change with broader application of the Phish Scale to a larger variety of phishing emails. A larger corpus of phishing emails will be needed to validate the cue count ranges.

It should be emphasized that the number of cues alone does not determine the detection difficulty for a target audience; it is only when considered in conjunction with the premise alignment that a detection difficulty rating can be computed.

TABLE III. PHISHING EXERCISE DATA

Exercise	**Number of cues**	**Premise alignment**	**Difficulty rating**	**Actual phishing click rate**
New voicemail (n = 69)	11 (Some)	Medium	Moderately difficult	11.6 % (8/69)
Unpaid invoice (n = 73)	8 (Few)	High	Very difficult	20.5 % (15/73)
Order confirmation (n = 66)	18 (Many)	Medium	Moderately difficult	9.1 % (6/66)
Gmail (n = 64)	7 (Few)	High	Very difficult	49.3 % (39/73)
Weblogs (n = 73)	14 (Some)	High	Very difficult	43.8 % (28/64)
Valentine (n = 4097)	13 (Some)	Low	Moderately/ Least difficult	11.0 % (549/4977)
Security token (n = 5024)	12 (Some)	Medium	Moderately difficult	8.7 % (439/5024)

The security token phish was the only exercise with a data entry component—after clicking the link users were taken to a webpage requesting their credentials. We report the data entry rates here rather than in Table III. For the security token phish, 24.4 % (107/439) of clickers entered data on the credential-harvesting webpage. However, this is only 2.1 % (107/5024) of the total number of employees who received the phishing email. Given that roughly 75 % of clickers did not enter data on the webpage, it seems that additional suspicion was triggered on this page, likely due to being asked for credentials. This is certainly in line with the fact that mandatory yearly security awareness training at NIST in the past has focused heavily on not sharing credentials.

b) Observations

Through these seven phishing exercises, we applied our Phish Scale to a variety of phishing attack types. This includes link-based attacks (New voicemail, Order confirmation, Gmail, Weblogs, and Valentine), an attachment attack (Unpaid invoice, which mimicked the real-world Locky ransomware attack), and a data entry or credential-harvesting attack (Security token). Now that we have used the Phish Scale to determine the detection difficulty rating for seven phishing exercises, there are a few observations we can make.

All of the exercises having a detection difficulty rating of *Very difficult* also have relatively high click rates (Unpaid invoice: 20.5 %, Gmail: 49.3 %, and Weblogs: 43.8 %). However, the Weblogs exercise has many more cues than the other two exercises, and at 14 cues, was at the extreme end of the *Some* cues range (9 to 14).

All of the exercises having a detection difficulty rating of *Moderately difficult* have relatively lower click rates (New voicemail: 11.6 %, Order confirmation: 9.1 %, and Security token: 8.7 %). The Valentine exercise has the detection difficulty range *Moderately difficult* to *Least difficult* and has a click rate of 11.0 %. The Valentine and Security token exercises have relatively larger and more varied sample sizes than the other exercises which likely makes it more difficult to categorize the premise alignment. And finally, we do not have an exercise with a detection difficulty rating of *Least difficult*, calling attention to the need to apply the Phish Scale to additional exercises.

c) Limitations

This work is an early effort to characterize phishing message detection difficulty for email users situated in their normal email processing environments. As such, the authors acknowledge there are certainly limitations with this work at this time.

Current notable limitations in this work include: 1) the list of cues is long but not exhaustive; 2) the uneven saliency of cues is not reflected; 3) categorizing premise alignment is not formulaic; 4) cue count ranges need to be informed by additional data; and, 5) additional data are needed for scale validation.

The authors anticipate that each of these limitations will be addressed as the Phish Scale is developed further.

V. Discussion and Future Directions

A. Click rates alone are insufficient: Why phishing detection difficulty matters

CISOs responsible for overseeing embedded phishing awareness training are often concerned when they observe click rates that are higher than expected. They are left wondering why click rates continue to be variable—possibly including large spikes—despite spending a significant amount of money and time training staff. CISOs must justify their cyber awareness training budgets and show a good ROI, lest their funding for such training be reduced. Unfortunately, if click rates continue to be high or variable, it is often—and we posit, incorrectly—perceived as due to ineffective training. We argue that this perception is fundamentally incorrect and hope to begin dispelling this perception through our development of a Phish Scale. Furthermore, we argue against focusing solely on phishing exercise click rates, and instead strongly encourage the inclusion of reporting rates and reporting times as well; these metrics must be considered in conjunction, not in isolation, as early reporting can greatly improve mitigation efforts. Are reporting rates higher than click rates? Is time to first report sooner than time to first click?

We hope to frame the discussion around high click rates in a way that makes sense to CISOs and argue that high click rates can indicate that users are being exposed to new, difficult, and contextually relevant phishing campaigns. We firmly believe difficult exercises actually improve user training effectiveness and awareness for real-world threats more than solely repeating the same or very similar, easier-to-detect phish. Click rates must be considered in conjunction with a deeper understanding of the phishing emails themselves and in light of reporting behavior as well. To this end, we have developed a Phish Scale to aid CISOs in better understanding and characterizing the detection difficulty of a given phishing exercise. Using operational data, the scale provides an indication of the difficulty email users in a target population will have detecting a particular phishing message. The Phish Scale addresses multiple components of phishing detection difficulty: cues, such as [26] and [30], and user context alignment [14]. Although our Phish Scale cue list is quite extensive, it is by no means exhaustive.

We expect that the three detection difficulty ratings we identified, *Very difficult*, *Moderately difficult*, and *Least difficult*, may eventually equate to click rate ranges. In speaking with CISOs, we anticipate ranges roughly along these lines: the *Very difficult* category having click rates above 20 %, the *Moderately difficult* category having click rates in the approximately 9 to 20 % range, and the *Least difficult* category having click rates below 9 %. We plan to inform the actual ranges with additional empirical data; the seven exercises presented here are a start.

It is early days for the Phish Scale, however, we believe the conceptual framework has promise when we consider the projected detection difficulty rating and the actual click rates for the seven exercises we examined. Additionally, we stress the Phish Scale components are still in development. We know all cues do not have equal salience. Finding an abbreviated method for CISOs to characterize premise alignment has proven difficult and elusive thus far. And finally, additional components such as severity of consequences and other factors may need to be considered sooner rather than later.

Indeed, we already see indications that additional factors warrant investigation for inclusion. For example, in the Weblogs exercise, there are 14 cues (categorized as *Some*), but this is right on the cusp of the *Many* cues category—and a corresponding easier detection difficulty rating. However, the actual click rate is very high, 43.8 % indicating detection is indeed difficult. The severity of the consequences in this premise is also very high—loss of job—suggesting its potentially strong influence.

B. Differential cue salience: Not all cues are created equally

Capturing the effect of phishing message cues is difficult, as not all cues are created equally. The saliency and effect of any particular phishing cue varies, determining whether it is perceived as a suspicion indicator versus a compelling hook. This aspect of phishing message characteristics is important to note. Whether a cue is perceived as a phish indicator versus a hook depends on the user and the user's context when processing the email. Well-known phishing indicators such as misspellings and grammar errors are often regarded by email users as suspicion-generating, and when noticed can lead to additional user scrutiny of the message for more phishing indicators. Another undisputed phishing characteristic is urgency. Its use is so common that it should be a red flag, however, urgency is legitimately common-place in today's world, diluting its suspicion-generating signal strength. Additionally, urgency inhibits System 2 processing [22] making it more a hook enhancer than a red flag.

C. *Categorizing user context and premise alignment*

In this initial version of a Phish Scale, we have used the terms *High*, *Medium*, and *Low* to bucket premise alignment for a target audience into intuitive high-level categories with associated definitions. While this is sufficient for the beginning phase of scale development, we may seek to refine the characterization methods for these categorical variables in future work, by investigating contextual relevance measures and scales. "Contextual" is a part of existing scales in other domains such as, "A Contextual Measure of Achievement Motivation" [35] and "contextual performance" as a dimension of individual work performance [24]. How might such existing scales and measures be leveraged for use in the phishing domain? Additionally, how do we account for changes in context over time?

Changes in contextual relevance may occur over quite long timescales, as someone slowly adds or changes job responsibilities over the years of their career, or very short timescales, as some event that day/week/month may trigger heightened contextual relevance. For example, Greene et al. [14] explained that users were concerned over a real-world vendor invoice that was unpaid, leading to temporarily heightened contextual relevance for the unpaid invoice phishing email. Daily events, such as expecting or missing a phone call, can temporarily heighten the contextual relevance of a "new voicemail" phishing email. Factors such as being busy, stressed, or rushed can also fluctuate widely during a work day. It is likely the case that there is a relatively fixed component of user context, in addition to a more time-sensitive, variable component. The current Phish Scale does not break down context and associated premise alignment into these subcomponents. It is unclear whether such a fine-grained distinction is indeed necessary at this point.

Although it may be quite feasible to discern premise alignment with finer granularity than our existing categories, this may actually be superfluous for the intended audience of the Phish Scale. With our goal of developing a simple, easy to use Phish Scale for CISOs and those responsible for implementing and overseeing phishing awareness training programs, it is likely the case that *High*, *Medium*, and *Low* categories for premise alignment are sufficient. The important point we seek to emphasize with our Phish Scale is that a highly relevant context makes it extremely difficult for users to detect phishing emails. The greater the contextual relevance, the less likely a user is to notice, attend to, and think deeply about suspicious email cues. Daily stressors such as time pressure in general reduce the cognitive resources that users have available to dedicate to email processing. When cognitive resources are reduced, it makes it more likely that users will engage in faster, heuristic, System 1 processing rather than thoughtful, slower, deeper System 2 processing [22].

A final point with respect to user context and premise alignment has to do with the size of the target audience: categorizing premise alignment becomes more difficult as the size of the target audience increases. With a larger target audience, there is typically a much greater variety of work responsibilities present and a wider variety of user contexts, which may or may not align with a phishing email premise.

D. *Comparing phishing data across sectors*

Although cross-exercise and cross-sector phishing comparisons are frequently made, and are indeed quite valuable, interpretation of such comparisons still pose significant challenges. In particular, when the level of phishing detection difficulty can vary so dramatically based on user context and premise alignment, it is in some sense a meaningless comparison without a basic understanding and assessment of: 1) characteristics of the phishing email itself and 2) characteristics of the target user population. More specifically, one must understand the premise and cues contained within a given phish in conjunction with the work context of the target user population. Toward this end, we believe our Phish Scale shows great promise as a tool to help frame data sharing on click rates and reporting rates across exercises, organizations, and sectors.

As we refine and mature this tool with input from the larger usable security community, we hope to move the Phish Scale out of the research community and into operational use. For instance, we believe that beyond providing benefits to CISOs and phishing training implementers, our Phish Scale could also provide significant value to joint organizations responsible for sharing cyber threat intelligence data. For example, the Federal Bureau of Investigation (FBI), has an InfraGard program, a partnership between the FBI and the private sector dedicated to sharing information and intelligence [10]. There are other such collaborative programs as well, for example, the National Cyber-Forensics and Training Alliance (NCFTA) is a nonprofit partnership between private industry, government, and academia working together to disrupt cybercrime [27]. Phishing in particular, and social engineering in general, are active threats across all industry verticals. By providing a phishing difficulty rating framework, our Phish Scale can help facilitate collaboration using a common language surrounding human phishing threat detection.

E. *Future work*

We encourage other usable security researchers and practitioners to use our Phish Scale, apply it to a much wider variety of phishing emails, and test its predictions against both existing phishing training exercise data, and ultimately against real-world phishing emails as well. We plan to continue applying our Phish Scale to a larger corpus of additional emails for which we have click rate and premise alignment data, and plan to partner with external entities to do the same. Unfortunately, our access to concurrent reporting data is more limited. A notable challenge of conducting research with operational workplace data is that there is often a tradeoff between experimental control and ecological validity. In this case, we had the benefit of extremely high ecological validity, as users were in their normal workplace settings with their normal tasks and email loads, but without the control necessary to capture reporting rates at the time of the phishing exercises. Nonetheless, the benefit of having new, in situ workplace data for ~ 5 000 employees offers an important contribution to the phishing literature and to the larger usable security community.

Beyond applying and testing the current Phish Scale with additional data, we intend to explore new scale components as well. For instance, it appears the Phish Scale would benefit from incorporating a measure of perceived consequence severity. Greene et al. [14] found that clickers were concerned over consequences arising from *not* clicking, such as failing to be responsive to their job duties. In contrast, non-clickers were more concerned over consequences due to clicking, such as accidentally downloading malware. Additionally, concern over consequences varied depending on the premise of the phish. In the phishing exercises described in the current paper, it is likely that concern over consequences was much higher for the Weblogs exercise, with its implied consequence of disciplinary action, to include dismissal, and the Security token exercise, which could have been concerning for teleworkers, as a security token is needed to access the organization's network from off campus. The current instantiation of the scale does not specifically address concern over consequences or the perceived relative severity of consequences. It may be possible that premise alignment alone is sufficient and already captures these effects for the Phish Scale's intended purpose, but additional research on this would be beneficial.

Additionally, we would like to investigate incorporating work on personality factors, and ultimately folding the various components of our Phish Scale into a lens model, an application of multiple regression often used in judgement and decision-making research. This would build upon prior lens modeling work by Tamborello and Greene [36] and Molinaro and Bolton [26]. Additional modeling and simulation research could explore the predicted click rates and reporting rates for different combinations of cues, context alignment, personality types, and phishing premises. How do different combinations affect phishing susceptibility? For example, consider this combination: users scoring high on conscientiousness, with a financial work context, who receive a phishing email with an authoritarian/time-sensitive transfer of funds premise, and very few suspicious cues. What if everything were the same but the work context, is that difference alone sufficient for someone to catch this phish? While we believe that context may trump all, additional research is necessary to see in which scenarios this holds, as well as, how and when it may change. One could simulate—with a well-validated model—the large number of possible combinations, to determine where to focus research and training intervention efforts based on quantified predicted risk metrics, such as the likelihood of clicking versus reporting.

F. Broader implications

In this section, we move beyond discussion of immediate future plans for the Phish Scale and into a discussion of broader implications for our work. The Phish Scale—and indeed phishing in general—is part of a much larger research agenda that addresses a spectrum of usable security issues. For instance, understanding risk, including human risk, is a key component of any organization's cybersecurity strategy, and risk management frameworks play an important role in helping maintain security and privacy [28]. Ultimately, we hope our Phish Scale can be used to help CISOs better understand and characterize their organization's phishing risk, by essentially profiling the types of phishing premises their users are more or less susceptible to as well as the organization's actual threats. Such data can be used to prioritize training efforts on more targeted interventions, and to prioritize investigative efforts for real-world suspected phishes. Targeted training interventions will likely need to move beyond embedded phishing exercises, especially for repeat clickers. In-person seminars, posters, informal lunch and learn sessions, and so on, are all part of a larger security awareness program. Additional interventions may include special email Graphical User Interface (GUI) elements or flagging, or perhaps more aggressive email filtering for certain users or groups based on their risks and job responsibilities.

In addition to risk profiling and targeted training, future work is also needed to understand how new technological email security measures will impact phishing. In particular, government agencies are quickly moving toward email authentication by implementing protocols such as Domain-based Message Authentication, Reporting, and Conformance (DMARC) and Domain Keys Identified Mail (DKIM) per the Department of Homeland Security (DHS) Binding Operational Directive 18-01 [6]. How will that affect the phishing space? On the other hand, pretexting is already gaining in popularity and will likely continue to do so, especially if new technological solutions prevent or threaten the success of certain more "traditional" phishing email scams. As advances in technological protections make some attacks less effective, or even one day obsolete, the attacks will not stop, but rather will transition and evolve in response. For instance, it seems likely that other out-of-band social engineering methods will continue to gain in popularity. Phishing is but one component of a much larger social engineering problem facing the cybersecurity field. Future work should examine how lessons learned in the phishing domain may inform other varieties of social engineering problems as well.

Acknowledgements

The authors gratefully acknowledge the institution's Information Technology Security and Networking Division and our Information Technology Security Officer partner for their support throughout this project, and thank the CISOs and training implementers who spoke with us regarding their phishing awareness programs, as well as anonymous reviewers for their comments.

Disclaimer

Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the products mentioned are necessarily the best available for the purpose.

REFERENCES

[1] M. Alsharnouby, F. Alaca, and S. Chiasson, “Why phishing still works: User strategies for combating phishing attacks.” International Journal of Human-Computer Studies, 2015, 82, pp. 69–82.

[2] M. Blythe, H. Petrie, and J.A. Clark, “F for fake: Four studies on how we fall for phish,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM. May 2011, pp. 3469–3478.

[3] C.I. Canfield, B. Fischhoff, A. Davis, “Quantifying phishing susceptibility for detection and behavior decisions.” Human Factors: The Journal of the Human Factors and Ergonomics Society, 2016, 58, pp. 1158–1172.

[4] D. Caputo, S.L. Pfleeger, J. Freeman, and M. Johnson, “Going spear phishing: Exploring embedded training and awareness,” in IEEE Security & Privacy, 12(1), January 2014, pp. 28–38.

[5] Cybersecurity Ventures, “2017 Cybercrime Report,” https://1c7fab3im83f5gqiow2qqs2k-wpengine.netdna-ssl.com/2015-wp/wp-content/uploads/2017/10/2017-Cybercrime-Report.pdf (Accessed Nov 2018).

[6] DHS, Department of Homeland Security. Binding Operational Directive BOD-18-01, 2017, https://cyber.dhs.gov/assets/report/bod-18-01.pdf (Accessed Nov 2018).

[7] R. Dhamija, J.D. Tygar, M. Hearst, “Why phishing works,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2006, pp. 581–590.

[8] J.S. Downs, M. Holbrook, and L.F. Cranor, “Decision strategies and susceptibility to phishing,” in Proceedings of the Second Symposium on Usable Privacy and Security (SOUPS ‘06), ACM, 2006, pp. 79–90.

[9] S. Egelman, L.F. Cranor, and J. Hong, “You’ve been warned: An empirical study of the effectiveness of web browser phishing warnings,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2008, pp. 1065–1074.

[10] FBI, Federal Bureau of Investigations. https://www.fbi.gov/about/partnerships/infragard (Accessed Nov 2018)

[11] B.J. Fogg, "Prominence-interpretation theory: Explaining how people assess credibility online." In CHI'03 Extended Abstracts on Human Factors in Computing Systems, ACM, 2003, pp. 722–723.

[12] S. Furnell, “Phishing: can we spot the signs?” Computer Fraud and Security 2007, pp. 10–15.

[13] S. Grazioli, “Where did they go wrong? An analysis of the failure of knowledgeable internet consumers to detect deception over the internet.” Group Decision and Negotiation, 2004, 13, pp. 149–172.

[14] K.K. Greene, M. Steves, M. Theofanos, and J. Kostick, “User Context: An Explanatory Variable in Phishing Susceptibility.” USEC NDSS 2018. Usable Security Workshop at the Network and Distributed Systems Security Symposium. February 18, 2018. San Diego, CA. DOI: https://dx.doi.org/10.14722/usec.2018.23016

[15] K.K. Greene, M. Steves, and M. Theofanos. “No Phishing beyond This Point,” in IEEE Computer, Cybertrust Column, 2018, vol. 51, no. 6, pp. 86–89. DOI: http://doi.ieeecomputersociety.org/10.1109/MC.2018.2701632

[16] Y. Han and Y. Shen. “Accurate spear phishing campaign attribution and early detection.” In Proceedings of the 31st Annual ACM Symposium on Applied Computing, ACM, 2016, pp. 2079–2086.

[17] Healthcare CyberGard Annual Conference, Charlotte, October 2018, https://www.ncfta.net/healthcare-cybergard-annual-conference-charlotte-october-2018/ (Accessed Jan 2019).

[18] Information Security and Privacy Advisory Board, https://csrc.nist.gov/CSRC/media/Projects/ISPAB/documents/minutes/ispab-june-2018-meeting-minutes.pdf (Accessed Jan 2019).

[19] M. Jakobsson, “The human factor in phishing.” Privacy and Security of Consumer Information, 2007, 7, pp. 1–19.

[20] M. Jakobsson and P. Finn, “Designing and conducting phishing experiments.” IEEE Technology and Society Magazine, Special Issue on Usability and Security, 2007.

[21] A. Karakasiliotis, S. Furnell, and M. Papadaki, “Assessing end-user awareness of social engineering and phishing.” In Australian Information Warfare and Security Conference, School of Computer and Information Science, 2006, Edith Cowan University, Perth, Western Australia.

[22] D. Kahneman, “Thinking, Fast and Slow.” New York: Farrar, Straus and Giroux, 2011.

[23] D. Kim, and J. Hyun Kim, "Understanding persuasive elements in phishing e-mails: A categorical content and semantic network analysis", Online Information Review, 2013, Vol. 37 Issue: 6, pp. 835–850, https://doi.org/10.1108/OIR-03-2012-0037

[24] L. Koopmans, C.M. Bernaards, V.H. Hildebrandt, H.C. de Vet, and A.J. van der Beek, “Measuring individual work performance: Identifying and selecting indicators.” *Work*, 2014, 48(2), pp. 229–238.

[25] D.E. Levari, D.T. Gilbert, T.D. Wilson, B. Sievers, D.M. Amodio, and T. Wheatley, “Prevalence-induced concept change in human judgement,” *Science* 29, June 2018: Vol. 360, Issue 6396, pp. 1465–1467. DOI: https://doi.org/10.1126/science.aap8731

[26] K.A. Molinaro, and M.L. Bolton, “Evaluating the applicability of the double system lens model to the analysis of phishing email judgments.” *Computers & Security*, 2018, 77, pp. 128–137.

[27] NCFTA, National Cyber-Forensics and Training Alliance. https://www.ncfta.net (Accessed Nov 2018).

[28] NIST, National Institute of Standards and Technology, 2018, “Special Publication 800-37, Revision 2, Risk Management Framework for Information Systems and Organizations—A System Life Cycle Approach for Security and Privacy”

[29] L. H. Newman, “What Spammers Could Do with Your Hacked Facebook Data,” in Wired, 2018, https://www.wired.com/story/facebook-hack-data-spammers/ (Accessed Nov 2018).

[30] K. Parsons, M. Butavicius, M. Pattinson, A. McCormac, D. Calic, and C. Jerram, “Do Users Focus on the Correct Cues to Differentiate Between Phishing and Genuine Emails?”, Australasian Conference on Information Systems 2015, https://arxiv.org/abs/1605.04717 (accessed Nov 2018).

[31] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, C. Jerram, “Phishing for the truth: A scenario-based experiment of users behavioural response to emails.” In IFIP International Information Security Conference, Springer, Berlin, Heidelberg, 2013, pp. 366–378.

[32] Ransomware, https://www.trendmicro.com/vinfo/us/security/definition/RANSOMWARE (Accessed Jan 2019).

[33] R. W. Rogers, “A protection motivation theory of fear appeals and attitude change.” J. Psychol, 1975, 91, pp. 93–114.

[34] B. D., Sawyer and P.A. Hancock, “Hacking the Human: The Prevalence Paradox in Cybersecurity.” Human Factors, 2018, 60(5), pp. 597–609.

[35] R.L. Smith, “A contextual measure of achievement motivation: Significance for research in counseling.” *VISITAS Online*, 2015.

[36] F.P. Tamborello, K.K. Greene. “Exploratory Lens Model of Decision-Making in a Potential Phishing Attack Scenario.” National Institute of Standards and Technology Interagency Report, NISTIR 8194, October 2017, DOI: https://doi.org/10.6028/NIST.IR.8194

[37] A. Tsow, M. Jakobsson, “Deceit and deception: A large user study of phishing.” Indiana University, Technical report TR649, 2007.

[38] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H.R. Rao. “Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model.” *Decision Support Systems*, vol. 51 no. 3, March 2011, pp. 576–586.

[39] A. Vishwanath, B. Harrison, Y.J. Ng, “Suspicion, cognition, and automaticity model of phishing susceptibility.” Communication Research, 2016, 0093650215627483.

[40] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H.R. Rao, “Phishing susceptibility: An investigation into the processing of a targeted spear phishing email,” in IEEE Transactions on Professional Communication, vol. 55, no. 4, December 2012, pp. 345–362.

[41] J. Wang, Y. Li, R. Rao, “Coping responses in phishing detection: an investigation of antecedents and consequences.” Inf. Syst. Res., 2017, 28, pp. 378–396.

[42] E.J. Williams, J. Hinds, and A.N. Joinson, “Exploring susceptibility to phishing in the workplace.” *International Journal of Human-Computer Studies*, 2018, 120, pp. 1–13.

[43] R. Wright, S. Chakraborty, A. Basoglu, and K. Marett, “Where did they go right?” Understanding the deception in phishing communications. Group Decision and Negotiation, 2010, 19, pp. 391–416.

APPENDIX A

The table below contains the list of operationalized cues, indicators and hooks we used to count the phishing message characteristics to obtain a rating of phishing detection difficulty. For each cue, the associated references and the criteria we used for counting are also provided.

TABLE: OPERATIONALIZED CUES

Cue Type	Cue Name	Description	References	Criteria for Counting
Error	Spelling and grammar irregularities	Spelling or grammar errors, mismatched plurality and so on	[2], [3], [8], [12], [13], [19], [20], [21], [31], [38], [40], [43]	Does the message contain spelling or grammar errors, including mismatched plurality?
	Inconsistency	Inconsistent content within the email	[14]	Are there inconsistencies contained in the email message?
Technical indicator	Attachment type	The presence of file attachments, especially an executable	[16]	Is there a potentially dangerous attachment?
	Sender display name and email address	Spoofed display names - hides the sender and reply-to email addresses	[2], [8], [12], [21], [23], [38], [39], [40]	Does a display name hide the real sender?
	URL hyperlinking	URL hyperlinking hides the true URL behind text; the text can also look like another link	[3], [8], [9], [12], [19], [21]	Is there text that hides the true URL in a hyperlink?
	Domain spoofing	Domain name used in email address and links looks similar to plausible	[14], [37]	Is a domain name used in addresses or links plausibly similar to a legitimate entity's domain?
Visual presentation indicator	No/minimal branding and logos	No or minimal branding and logos	[2], [12], [13], [19], [21], [23], [37], [39]	Is appropriate branding missing?
	Logo imitation or out-of-date branding/logos	Spoof or imitation of logo/out-of-date logo	[14]	Do any branding elements appear to be an imitation or out-of-date?
	Unprofessional looking design or formatting	Formatting and design elements that do not appear to have been professionally generated	[7], [11], [19], [21], [31], [37]	Does the design and formatting violate any conventional professional practices?
	Security indicators and icons	Security indicators and icons	[7], [19]	Are any inappropriate security indicators or icons present?
Language and content	Legal language/copyright info/disclaimers	Any legal type language such as copyright information, disclaimers, tax implications	[19]	Does the message contain any legal type language such as copyright information, disclaimers, tax information?
	Distracting detail	Distracting Detail	[14]	Does the message contain any detailed aspects that are not central to the content?
	Requests for sensitive information	Requests for sensitive information, like a Social Security number or other identifying information	[8], [12], [14]	Does the message contain a request for any sensitive information, including personally identifying information or credentials?
	Sense of urgency	Use of time pressure to try to get users to quickly comply with the request	[1], [3], [8], [12], [21], [23], [38]	Does the message contain time pressure, including implied?
	Threatening language	Use of threats such as legal ramifications	[1], [3], [8], [21], [23], [38]	Does the message contain a threat, including an implied threat?
	Generic greeting	A generic greeting and an overall lack of personalization in the email	[1], [3], [8], [9], [21], [31], [37]	Does the message lack a greeting or lack personalization in the message?
	Lack of signer details	Emails including few details about the sender, such as contact information	[23]	Does the message lack detail about the sender, such as contact information?
Common tactic	Humanitarian appeals	Appeals to help others in need	[21], [23]	Does the message make an appeal to help others?

	Too good to be true offers	Contest winnings or other unlikely monetary and/or material offerings	[13], [21], [31], [43]	Does the message offer anything that is too good to be true, such having won a contest, lottery, free vacation and so on?
	You’re special	Just for you offering… such as a valentine e-card from a secret admirer		Does the message offer anything just for you?
	Limited time offer	This offer won't last long…		Does the message offer anything for a limited time?
	Mimics a work or business process	Mimics any plausible work process such as new voicemail, package delivery, order confirmation, notice of invoice, and so on		Does the message appear to be a work or business-related process?
	Poses as friend, colleague, supervisor, authority figure	Email purporting to be from a friend, colleague, boss or other authority figure	[42]	Does the message appear to be from a friend, colleague, boss or other authority entity?

APPENDIX B

The table below contains the observed counts of each cue for each exercise we evaluated for this effort to-date.

TABLE: EXERCISE CUE COUNTS

		Exercise Cues Observed with Counts						
		11	8	18	7	14	13	12
Cue Type	***Cue Name***	**New Voicemail**	**Unpaid Invoice**	**Order Confirmation**	**Gmail**	**Weblogs**	**Valentine**	**Token**
Error	Spelling and grammar irregularities	1	1	2				1
	Inconsistency		1					1
Technical indicator	Attachment type		1					
	Sender display name and email address	1	1	1	1	1	1	1
	URL hyperlinking	1		6	1	1	3	1
	Domain spoofing	1	1	1	1	1	1	1
Visual presentation indicator	No/minimal branding and logos	1		1			1	1
	Logo imitation or out-of-date branding/logos							
	Unprofessional design or formatting	1					1	
	Security indicators and icons							
Language and content	Legal language/copyright info/disclaimers	1		1		1	1	1
	Distracting detail	2		2		1	1	
	Requests for sensitive information					1		1
	Sense of urgency		1	1	1	1	1	1
	Threatening language					3		
	Generic greeting			1	1	1	1	1
	Lack of signer details	1		1		1	1	1
Common tactic	Humanitarian appeals							
	Too good to be true offers							

	You’re special						1	
	Limited time offer							
	Mimics a work or business process	1	1	1	1	1		1
	Poses as friend, colleague, supervisor, authority figure		1		1	1		