Behavior Research Methods (2021) 53:1342–1352
https://doi.org/10.3758/s13428-020-01495-0

The Phishing Email Suspicion Test (PEST) a lab-based task for evaluating the cognitive mechanisms of phishing detection

Ziad M. Hakim[1,2] · Natalie C. Ebner[2,3,4,5] · Daniela S. Oliveira[6] · Sarah J. Getz[5,7] · Bonnie E. Levin[5,7] · Tian Lin[2] · Kaitlin Lloyd[1] · Vicky T. Lai[1,8] · Matthew D. Grilli[1,5] · Robert C. Wilson[1,5,8]

Accepted: 30 September 2020 / Published online: 19 October 2020

Abstract

Phishing emails constitute a major problem, linked to fraud and exploitation as well as subsequent negative health outcomes including depression and suicide. Because of their sheer volume, and because phishing emails are designed to deceive, purely technological solutions can only go so far, leaving human judgment as the last line of defense. However, because it is difficult to phish people in the lab, little is known about the cognitive and neural mechanisms underlying phishing susceptibility. There is therefore a critical need to develop an ecologically valid lab-based measure of phishing susceptibility that will allow evaluation of the cognitive mechanisms involved in phishing detection. Here we present such a measure based on a task, the Phishing Email Suspicion Test (PEST), and a cognitive model to quantify behavior. In PEST, participants rate a series of phishing and non-phishing emails according to their level of suspicion. By comparing suspicion scores for each email to its real-world efficacy, we find initial support for the ecological validity of PEST – phishing emails that were more effective in the real world were more effective at deceiving people in the lab. In the proposed computational model, we quantify behavior in terms of participants' overall level of suspicion of emails, their ability to distinguish phishing from non-phishing emails, and the extent to which emails from the recent past bias their current decision. Together, our task and model provide a framework for studying the cognitive neuroscience of phishing detection.

Keywords Phishing · Cybersecurity · Decision making · Sequential effects

Introduction

Email phishing is a type of cyber social engineering attack in which seemingly legitimate emails attempt to lure the receiver into performing an action with negative consequences (e.g., opening a malicious attachment that installs malware on the victims' device). While the popular conception of phishing is a message from the infamous "Nigerian Prince," modern phishing emails can be hard to distinguish from safe emails, with large-scale studies suggesting click rates as high as 20% for the most effective phishing emails (PhishMe, 2016; Williams, Hinds, & Joinson, 2018). In part because of this high click rate, phishing is estimated to cost tens of billions of dollars each year (Smart People Easier, 2014) and is now also recognized as a major public health problem associated with negative health outcomes that include depression and suicide (Button, Lewis, & Tapley, 2014; Fraud Advisory Panel, 2015).

While technological solutions, such as filters and blacklists ("Google Safe Browsing," n.d.), massively reduce the number of phishing emails reaching people's inboxes, purely technical

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-020-01495-0.

✉ Robert C. Wilson
bob@arizona.edu

[1] Department of Psychology, University of Arizona, Tucson, AZ, USA

[2] Department of Psychology, University of Florida, Gainesville, FL, USA

[3] Department of Aging and Geriatric Research, Institute on Aging, University of Florida, Gainesville, FL, USA

[4] Florida Institute for Cybersecurity, University of Florida, Gainesville, FL, USA

[5] Evelyn F. McKnight Brain Institute, Gainesville, FL, USA

[6] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA

[7] Department of Neurology, Miller School of Medicine, University of Miami, Coral Gables, FL, USA

[8] Cognitive Science Program, University of Arizona, Tucson, AZ, USA

defense solutions can never be perfect. This is in part because of the sheer volume of phishing attempts, which are estimated to account for 1 in every 392 emails ("Overview of fraud and computer misuse statistics for England and Wales - Office for National Statistics," n.d.), and because of the sophistication of many phishing emails, which are carefully crafted to avoid filters. Further, phishing messages and corresponding malicious landing pages change constantly (Google, n.d.), adding challenges to machine learning methods that try to filter unwanted messages. Indeed, there are even websites (such as ("Test and Optimize your Emails for the Inbox," n.d.)) where attackers can test whether a particular email will be flagged or not by several leading email providers. Thus, human decision-making is the last line of defense against phishing and there is much interest in understanding the cognitive and neural processes underlying phishing detection in order to identify those individuals who are most susceptible (Ebner et al., 2020; Lin et al., 2019; Oliveira et al., 2017) as well as to evaluate the outcome of training programs designed to reduce phishing victimization (Norris, Brookes, & Dowell, 2019).

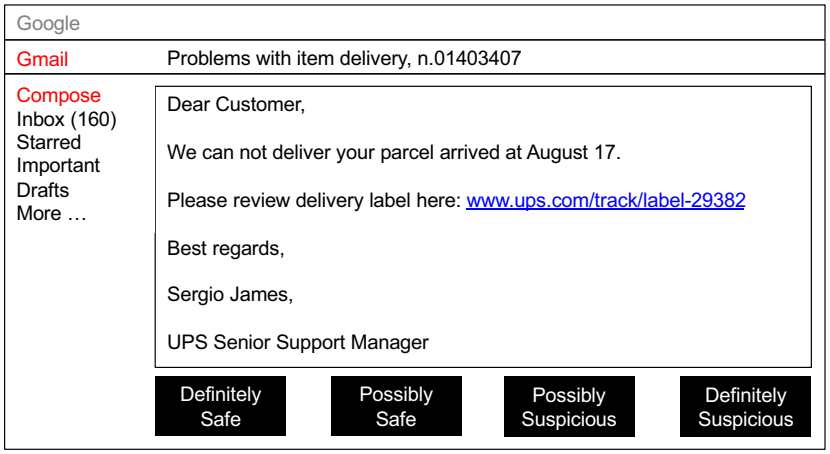
Perhaps the gold standard for measuring an individual's susceptibility to phishing deception is to actually try to phish them in the home or office setting – that is, to send them a phishing email and measure whether they engage with it or not (Caputo, Pfleeger, Freeman, & Eric Johnson, 2014; Kumaraguru et al., 2009; Lin et al., 2019; Luo, Zhang, Burd, & Seazzu, 2013; Oliveira et al., 2017; Vishwanath, 2015; Williams et al., 2018). Crucially, in these studies, participants are not aware that they have signed up for a phishing study and are not primed to be expecting suspicious emails (although in some studies, participants did consent to installing a browser plugin, which they were told was necessary to track their daily Internet use; Oliveira et al. 2017; Lin et al. 2019). Therefore, this direct approach has high ecological validity because the only difference from a genuine phishing attack is that the participants are not harmed. On the flip side, however, because the experiment occurs outside of the lab, these "field" experiments are less controlled than a lab study and any "in-the-moment" correlates of phishing susceptibility (such as physiological and neural measures) are almost impossible to obtain. Another challenge is that the number of emails sent to participants in field experiments is often small because of the need to simulate real-world phishing (but see Lin et al. 2019), where emails might be spaced out over several days to avoid arousing suspicion. This approach results in limited sampling of each person's susceptibility to attack, and this small amount of data per person makes it difficult to test cognitive models of how individual phishing emails are evaluated and detected.

Another approach is to try to measure phishing susceptibility in the lab. In the most direct version of this approach, the experimenter tries to phish people in the lab, for example by having them browse safe and phishing websites to see whether or not they divulge sensitive information (Gavett et al., 2017). Other lab-based experiments involve roleplaying another person checking their emails (Downs, Holbrook, & Cranor, 2007; Sheng, Holbrook, Kumaraguru, Cranor, & Downs, 2010) or rating a series of emails according to how suspicious they appear or how likely the person would be to respond to such an email (Alsharnouby, Alaca, & Chiasson, 2015; Dhamija, Tygar, & Hearst, 2006; Jones, Towse, Race, & Harrison, 2019; Kelley & Bertenthal, 2016; Rajivan & Gonzalez, 2018; Wood, Liu, Hanoch, Xi, & Klapatch, 2018; Yan & Gozu, 2012). Despite the increased experimental control offered by these lab-based tasks, the extent to which these measures are ecologically valid remains unknown. Of particular concern is that, in many of these studies, participants are told that some emails will be suspicious, thus priming them to expect phishing emails, which may cause them to engage different mental processes for email evaluation, than they would use when evaluating emails in the real world. Given the urgency and magnitude of this growing crisis from phishing on health and well-being, there is a need to document and quantify the relationship between lab and real-world behavior.

In this paper, we provide initial evidence that lab and real-world phishing susceptibility are related. Using a new task, the Phishing Email Suspicion Test (PEST), we asked participants to rate the degree to which a series of 160 emails appeared suspicious. By design, 84 of these emails (the "simulated-phishing" emails see Fig. 1) had been previously used in a field-experimental phishing study using an independent sample of participants (Lin et al. 2019; Oliveira et al., 2017). This allowed us to determine whether emails that were effective in real life (i.e., people had fallen for them in the field experiment in which they were not aware that they were phished), were also effective in the lab (i.e., were rated as *less* suspicious in PEST when participants were explicitly told to evaluate the suspiciousness of the emails). With this design, which used different sets of participants in the lab- and field-based components of the study, we cannot assess whether PEST behavior correlates with real-world phishing susceptibility of the individual. Despite this limitation on the interpretation, this initial test of ecological validity is important. If there is no correlation between in-lab and out-of-lab behavior across emails, then the relevance of PEST – as well as possibly other lab-based measures of phishing susceptibility – as a measure of real-world behavior is in question.

In addition to testing the ecological validity of phishing detection in the lab, in this paper we aimed to investigate the cognitive mechanisms of PEST behavior using a computational model. This model was inspired by the similarity between PEST, in which participants evaluate a series of emails according to their level of suspicion, and more traditional tasks from psychophysics, in which participants evaluate a series of physical stimuli (e.g., loudness of sounds or heaviness of weights) according to their perceived properties. The main difference between PEST and classical psychophysics

a) Task display with real phishing email



b) Text from example emails

SIMULATED PHISH	REAL SAFE	SIMULATED SAFE
Free Gift Received From Matt Dear Bob, Your friend Matt sent you a free gift through Games with Friends, a new social media site for keeping in touch through games and trades. Check out what Matt got you by following the link below: http://www.tucsonweekend.com/gift-social/ Matt Games with Friends	Please confirm your e-mail address - Airbnb Hi Bob, Welcome to Airbnb! In order to get started, you need to confirm your email address. Confirm Email Thanks, The Airbnb Team Sent with Love from Airbnb Airbnb, Inc., 888 Brannan St, San Francisco, CA 94103	Apple iCloud Security Dear Bob, Your AppleID was used to sign in to iCloud on an iPhone 6S in Tucson, Arizona at 7:06 PM on Monday, January 8, 2018. If this was your sign-in, please disregard this email. If this was not your sign in, please contact Apple Support at the link below. https://support.apple.com iCloud Authentication Services Copyright 2018 Apple Inc. All Rights Reserved

Fig. 1 The Phishing Email Suspicion Test (PEST). Participants were presented with a series of 160 emails from one of four categories (real-phish, simulated-phish, real-safe, simulated-safe; 40 emails per category) in randomized order. Participants were instructed to rate each email on a four-point scale from "definitely safe" to "definitely suspicious." (a) Schematic of the display seen by participants when evaluating an email. In this case, the email is a real-phishing email. (b) Examples of the subject line and text from the other types of email: a simulated-phishing email, a real-safe email, and a simulated-safe email. Note that in the lab-based PEST, participants saw all four email types, while in the field-based PHIT, participants were only sent emails from the simulated-phishing category

tasks is that, unlike loudness or weight, there is no *objective* measure of suspiciousness. Instead we need to norm the stimuli, using the judgments of others to compute an average suspicion score for each email. We then treat these subjective measures of suspiciousness just like the objective measures of loudness or weight. This allows us to quantify behavior in terms of a slope, the extent to which each person's rating correlates with the mean suspicion score of the group, and a bias, their propensity to say that any email is suspicious. Crucially, and unlike overall accuracy on the task, these measures are largely independent of the emails used in the task, allowing us to estimate individual difference parameters (i.e., slope and bias) that should predict behavior on any set of emails (so long as these emails are appropriately normed).

In addition to quantifying the slope and bias, our model also allows us to investigate the effect of previously observed emails on the current judgment. Such sequential effects have a rich history in psychophysics and there are well-established models as to how they might arise. One model in particular (Ward & Lockhead, 1970) suggests that people assign a rating to a stimulus by comparing it with previous stimuli. For example, in the case of loudness rating, participants compare the loudness of the current stimulus to that of the previous stimulus. If they judge the current stimulus to be louder, they give it a higher rating; if they judge the current stimulus to be quieter, they give it a lower rating. Such a comparison process reveals itself in the form of sequential effects in the ratings, whereby the current rating is biased by an "assimilative" effect of the

past rating (i.e., the current rating is biased towards the past rating), and a "contrastive" effect of the past stimulus (i.e., the current rating is biased away from the past stimuli) (Ward & Lockhead, 1970). Thus, if email evaluation involves a similar comparison process with past emails, then it should reveal itself in the sequential effects. To test this idea, we included terms for the past rating and past stimulus (i.e. normed suspicion score of the last email) in the model. This allowed us to ask whether there was a positive (assimilative) effect of past rating and a negative (contrastive) effect of past stimulus.

Methods

Code and data availability

Code to run the task is available on GitHub at https://github.com/zmhakim/PEST. Data and Matlab code to reproduce all figures in this paper is available on GitHub at https://github.com/bobUA/HakimEtAl_PEST_dataAndCode. The PEST data is also available on dataverse at (Hakim, et al. 2020).

Participants

Ninety-seven undergraduates (36 male, 61 female, ages 18–27, mean 18.9, standard deviation 1.39) were recruited from the University of Arizona Psychology subject pool. An additional three participants completed the task as part of their enrollment in the subject pool but were excluded from the analysis because they were under 18 (as required by our subject pool-IRB agreement). This sample size was determined as part of a larger study to detect age-related differences in phishing. This power analysis suggested we would need 80 participants in each age group (younger and older) to detect small to medium sized age effects in an intervention-based study (control vs. intervention). We opted to run more younger participants in this initial study of PEST to account for participants under 18 years of age, as well as non-compliance that often occurs in subject-pool experiments. Although we did not perform a formal power analysis to determine the sample size required for the analyses performed in this paper, we reasoned that this relatively large sample size (at least compared to in-lab studies of phishing susceptibility) would be sufficient to assess the basic differences in suspicion score between safe and phishing emails. Participants were not paid for participation but received course credit for completing the experiment. The study was approved by the Institutional Review Board at the University of Arizona and all participants provided written informed consent.

The Phishing Email Suspicion Test (PEST)

Participants were presented with a series of 160 emails and instructed to classify on a four-point scale from 1, "definitely safe," to 4, "definitely suspicious" via keyboard press (see Fig. 1 for schematic display and sample emails). Participants were told to maximize their classification accuracy in order to maximize their "score," although they received no monetary or other reward for their performance. The order of emails seen in the task was randomized for each participant. To minimize learning during the task, participants did not receive feedback about their performance until the end of the experiment. The task was coded using Psychtoolbox for Matlab (Brainard, 1997). Key press as well as response times were recorded.

Emails varied in their source (real emails vs. simulated emails created by us) and legitimacy (safe vs. phishing) such that there were four types of email (real-safe, simulated-safe, real-phish, and simulated-phish). Each participant saw 40 examples of each type of email, but because we had more than 40 examples of some email types, they did not all see the same emails. Overall, the 160 emails seen by participants were sampled from 348 emails made up of 140 real-phishing emails, 84 simulated-phishing emails, 84 simulated real emails and 40 real emails.

Real-phishing emails were sampled from a set of one hundred and forty genuine phishing emails obtained from four websites that collect phishing examples ("Information Security at UVA," n.d., "Phishing and Scam Emails: A Realtime Database of Phishing Emails," n.d.; Information Security University of Arizona, n.d.; UCLA Information Security Office, n.d.). These emails were edited to improve relevance for University of Arizona students only in cases where they mentioned specific institutions or locations (e.g., University of Virginia was changed to University of Arizona). The real-phishing emails were selected such that they spanned a range of effectiveness, from obvious phishing to more convincing attacks (based on subjective judgments of two of the authors, ZMH and KL). Simulated-phishing emails were taken from the 84 emails used in the PHishing Internet Task (PHIT) (Lin et al., 2019; Oliveira et al., 2017), see below for details.

Safe emails also comprised real and simulated messages. Forty real-safe emails were taken from the authors' personal email inboxes and included messages from banks and PayPal, as well as password resets or general account management. Eighty-four simulated-safe emails were adapted versions of the original PHIT emails that were altered to make them seem less suspicious, but matched on other features such as word count, purported identity of the sender and topic. In particular we corrected for poor syntax, grammatical errors, informal language, and aggressive tone in the original PHIT emails. Because the simulated-safe emails were created by editing the simulated-phishing emails from the original PHIT study

(Lin et al., 2019; Oliveira et al., 2017), the simulated-safe and simulated-phishing emails were almost perfectly matched for length, topic, and the requested action (click a link, download an attachment). We also attempted to match the real-safe and real-phishing emails to the simulated emails on the same parameters.

Finally, because there were more than 40 of each of the real-phishing, simulated-phishing, and simulated-safe emails, not all participants saw all of these emails (see Supplementary Fig. 1A for a plot showing the number of people encountering each email). We sampled the specific emails participants saw at random, which raises the concern that the distribution of suspicion scores, and thereby difficulty of the task, could have varied from person to person (e.g., one person only sees obvious phishing emails, while another only sees the most convincing phishing emails). However, a follow-up analysis of the distribution of suspicion scores of emails that participants actually saw (i.e., the average rating for each email) alleviated this concern by showing that the distribution of the suspicion score across emails, via random sampling, was similar across participants (Supplementary Fig. 1B).

Real-world phishing efficacy for a subset of emails

Phishing efficacy of the 84 simulated-phishing emails had previously been assessed in a field experiment by our group (Lin et al., 2019; Oliveira et al., 2017). In this Phishing Internet Task (PHIT), 158 participants were sent emails to the email address they had registered with the study, and a browser plugin recorded whether participants clicked on the link embedded in each email. Crucially, in this study participants were deceived about the nature of the study and were not expecting us to try to phish them. Thus, the behavioral data on PHIT allowed us to construct a measure of phishing efficacy for each email had it occurred in real life, as the proportion of times an email was successful in eliciting a click. In particular, each participant received 21 emails each during the course of 21 days (one per day). Emails had been created based on a large set of authentic spam emails from an independent sample of Internet users (Lin et al. 2019; Oliveira et al., 2017). On average each email was sent to just under 40 participants (range 37–42).

Regression model of PEST behavior

To analyze PEST behavior, we built a linear regression model similar to that used in (Wilson, 2018). This model assumes that participants' ratings for each email are computed according to

$$c_t = \beta_{bias} + \beta_f f_t + \beta_{fPast} f_{t-1} + \beta_{cPast} c_{t-1} \quad (1)$$

where c_t is the rating on trial t, f_t is the current stimulus "strength," computed as the average suspicion score across participants for this stimulus, f_{t-1} is the previous stimulus, and c_{t-1} is the previous choice. The regression weights are the free parameters of the model and denote the overall phish bias (β_{bias}), the effect of the current stimulus (β_f), the effect of the last stimulus (β_{fPast}), and the effect of the last rating (β_{cPast}). For the purposes of the regression (and to keep the notation in line with our previous work; Wilson, 2018), the ratings and suspicion scores were normalized to range between -1 and +1. Regression weights were estimated separately for each participant using a standard least-squares method (implemented via the `glmfit` function in Matlab).

Results

Participants were more suspicious of phishing than safe emails

Overall, participants performed above chance in PEST, achieving a mean accuracy of 62% (defined as the safe and phishing emails that were correctly classified as possibly or definitely safe or phishing). Regarding mean suspicion scores for each of the four email types (real-safe, simulated-safe, real-phish, simulated-phish), we conducted a repeated measures ANOVA with email legitimacy (safe/phish) and source (real/simulated) as within-subject factors (Supplementary Table 1). This analysis revealed a main effect of legitimacy ($F(1, 288) = 455.1$, $p < 0.001$, partial $\eta^2 = 0.61$), a main effect of source ($F(1, 288) = 56.3$, $p < 0.001$, partial $\eta^2 = 0.16$), and an interaction effect between legitimacy and source ($F(1, 285) = 66.5$, $p < 0.001$, partial $\eta^2 = 0.19$). Additional analysis further showed that these effects were due to participants rating phishing emails as more suspicious than safe emails ($t(96) = 20.3$, $p < 0.001$), simulated-safe emails as more suspicious than real-safe emails ($t(96) = 11.3$, $p < 0.001$), and simulated-phishing emails the same as real-phishing emails ($t(96) = 0.48$, $p = 0.63$) (Fig. 2a). Despite this difference in overall suspicion score between real and simulated-safe emails, ratings for both types of simulated emails were highly correlated with ratings of real emails (Fig. 2b, for safe emails $r(95) = 0.56$, $p < 0.001$; Fig. 2c for phishing emails $r(95) = 0.45$, $p < 0.001$).

Item analysis revealed a wide range of suspicion scores across emails

Averaging across participants allowed us to compute a suspicion score for each email. As shown in Fig. 3, there was a wide range of suspicion scores for all four types of email. Note that this analysis is different than the analysis presented in Fig. 2, in which we averaged over emails to get suspicion scores for each person (i.e., participants as unit of analysis). In Fig. 3, we average over *people* to get a suspicion score for each email (i.e., emails as unit of analysis).

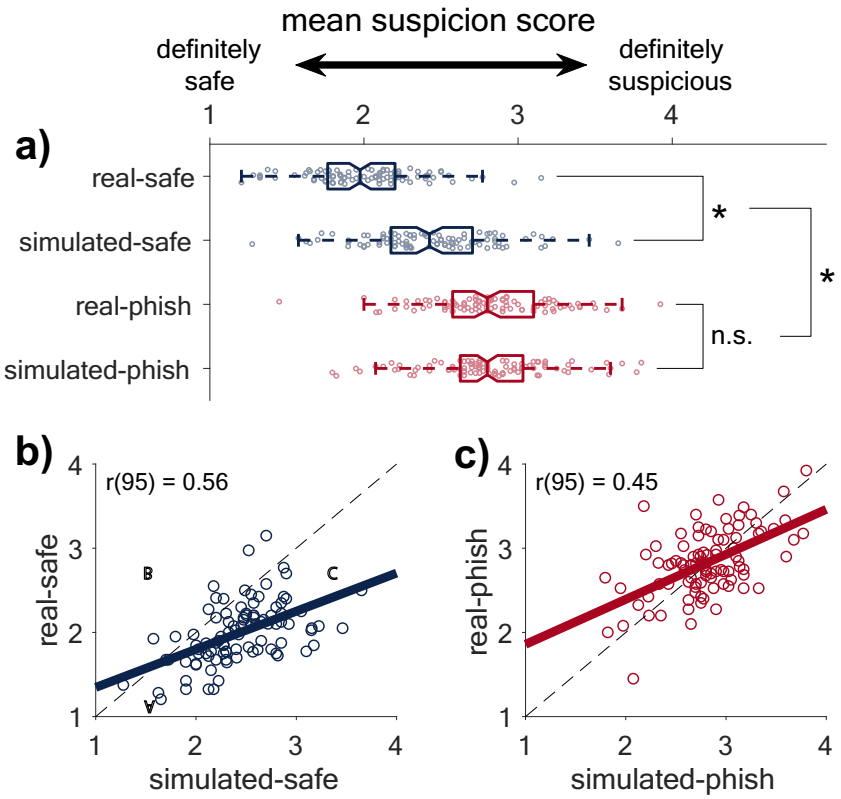


Fig. 2 Average suspicion scores for each participant in PEST for each email type. Safe emails are depicted in *blue* and phishing emails are depicted in *red*. Each *dot* represents the mean suspicion score given to each type of email by one participant. (a) Mean suspicion score for each of the four email types (real-safe, simulated-safe, real-phish, simulated-phish). Phishing emails were rated as more suspicious than safe emails and simulated-safe emails were rated as more suspicious than real-safe emails. *Notched box plots* represent median, confidence intervals for the median, upper and lower quartiles of suspicion scores and range. * indicates $p < 0.001$. (b, c) Correlation between suspicion scores for real and simulated emails. Each participant corresponds to a single dot. The correlations between simulated and real-safe (b) and simulated and real-phishing (c) emails suggest that performance on the simulated emails was positively associated with performance on real emails, for both safe and phishing emails

While safe emails were generally rated as less suspicious than phishing emails, there was considerable overlap in the ratings for the four types of email. For example, the least suspicious phishing email (mean score = 1.46) was ranked almost as safe as the least suspicious safe email (mean score = 1.38). Likewise, the most suspicious safe email (mean score = 3.26) was rated more suspicious than the average phishing email (mean score = 2.79). This overlap in suspicion score between phishing and safe emails illustrates how convincing some phishing emails can be.

More effective phishing emails were rated as less suspicious

As a test of the ecological validity of our newly developed PEST, we compared the average suspicion score of each email on PEST to the efficacy of the email as previously assessed in the original PHIT field experiment (Lin et al., 2019; Oliveira et al., 2017). Crucially, in the field-based PHIT, participants were not told that we would be trying to phish them meaning that the efficacy of an email in PHIT likely approximated the real-world efficacy of that email had it occurred in a real phishing campaign.

By design, 84 emails (the simulated-phishing emails) for which we had real-world measures of efficacy were included in PEST. As shown in Fig. 4a, we found a negative correlation between PEST and real-world behavior (Spearman's rho(82) = – 0.22, $p = 0.048$) such that emails with a low suspicion score in the lab (PEST) had been more effective at phishing participants in the real world field experiment (PHIT). As a complement to the correlation analysis, we ran a separate analysis in which we split the emails into two groups depending on whether they had been clicked on at all ("clicked", $n = 44$ emails) or not ("not-clicked", $n = 40$ emails) in PHIT. Consistent with the correlation analysis, clicked emails were rated as less suspicious than not-clicked emails ($t(82) = 2.09$, $p = 0.04$) (Fig. 4b). Together these findings provide initial evidence that PEST may be an ecologically valid measure of phishing behavior in the real world.

Regression model supported a contrastive effect of past stimulus and an assimilative effect of past rating

The regression model (see Methods) assumes that participants' ratings for each email are computed according to

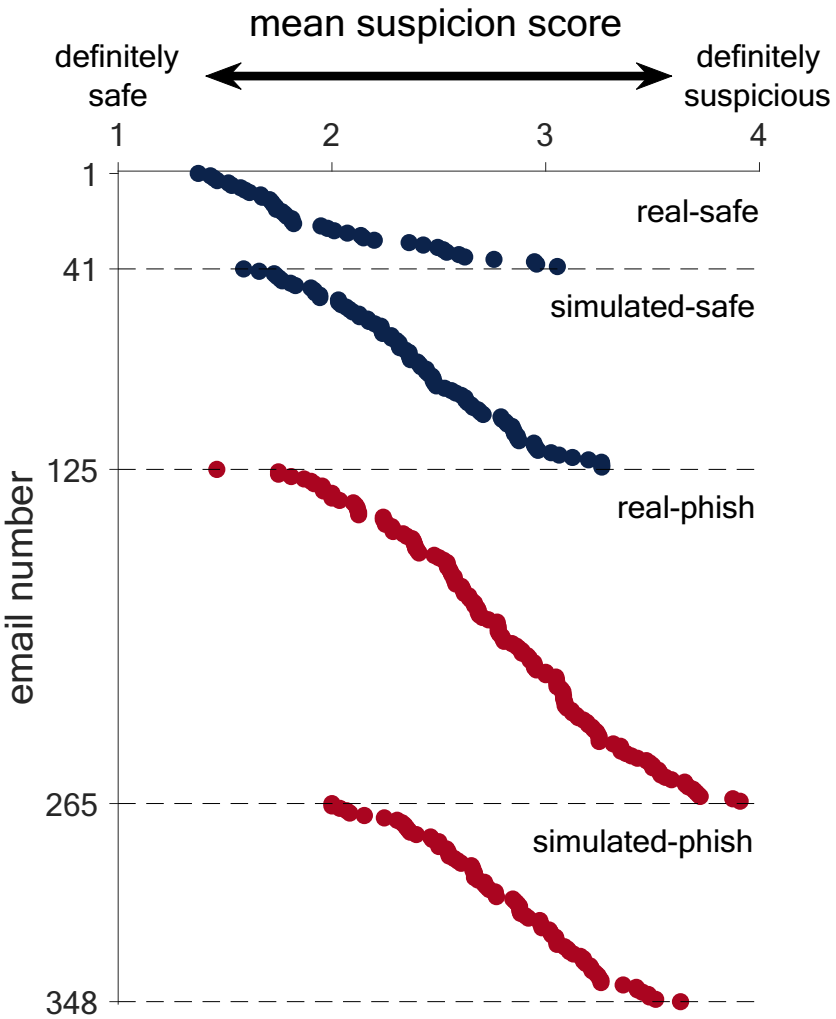


Fig. 3 Average suspicion scores for each email in PEST. Each email corresponds to a different dot. Emails were grouped by type (real-safe, simulated-safe, real-phish, simulated phish). Safe emails are depicted in *blue* and phishing emails in *red*. While, on average, safe emails were ranked as less suspicious, there was considerable overlap in the mean suspicion score for each type of email

$$c_t = \beta_{bias} + \beta_f f_t + \beta_{fPast} f_{t-1} + \beta_{cPast} c_{t-1} \quad (2)$$

where c_t is the rating on trial t, f_t is normed suspiciousness score of the current stimulus, f_{t-1} is the score of the previous stimulus, and c_{t-1} is the participant's previous rating. Behavior in this model is quantified in terms of the intercept (β_{bias}), the overall bias towards saying an email is suspicious, the slope (β_f), the extent to which the current stimulus drives the rating, and the sequential effects (β_{fPast}), capturing the effect of normed suspicion score of the last email, and the effect of the participants past rating (β_{cPast}).

Fit values of the regression weights are shown in Fig. 5. The largest effect was for the current stimulus, the slope, with a median regression weight close to 1, consistent with participants' score being correlated with the average score of other people, but with considerable variability across people, consistent with individual differences in phishing susceptibility. This effect of the current stimulus was positive, assimilative, for all participants and statistically significant ($t(96) = 27.28$, $p < 0.001$). There was also a significant phish bias. That is, all else being equal, participants were more likely to say that an email was suspicious than safe ($t(96) = 4.34$, $p < 0.001$). More subtly, we also saw evidence for sequential effects in the form of a negative, contrastive, effect of the last stimulus ($t(96) = -3.53$, $p < 0.001$) and a positive, assimilative, effect of the last rating ($t(96) = 4.40$, $p < 0.001$). These effects are consistent with previous work in sequential judgment tasks with no feedback and suggest a comparison process for email evaluation in which the current email is evaluated relative to the last (Holland & Lockhead, 1968).

Discussion

In this paper, we introduced PEST as a tool for assessing email phishing susceptibility in the lab with ecological validity, as

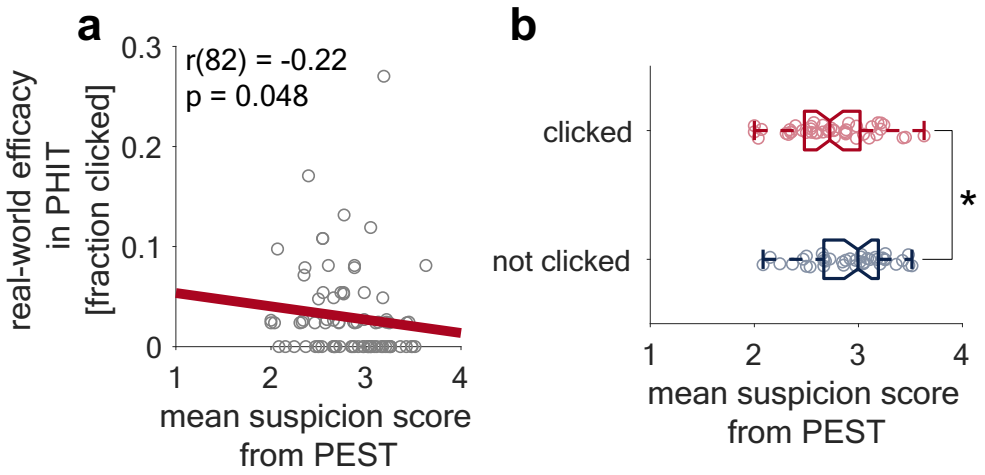


Fig. 4 Ecological validity of PEST. (a) Correlation between PEST behavior and real-world efficacy of each email in the PHIT field experiment, operationalized as the fraction of times the link in the email was clicked. Emails rated as more suspicious in PEST had been more effective at phishing people in the real-world PHIT paradigm. (b) Alternate analysis separating emails into two groups based on whether they had been successful at phishing at least one Internet user in the PHIT field experiment (i.e., at least one person had clicked on the link embedded in the email; "clicked" email) or not (no person had clicked on the link embedded in the email; "not-clicked" email). In line with the correlation analysis, emails that were not clicked in the original PHIT were also rated as more suspicious in PEST. * indicates $p < 0.05$

well as a cognitive model to quantify PEST behavior. In this task, participants evaluated a series of 160 emails regarding their level of suspicion. While participants were generally more suspicious of phishing than safe emails, their performance was far from perfect, with participants only classifying emails correctly about 62% of the time. Such a relatively low accuracy is consistent with previous reports, such as (Jones et al., 2019) who found an accuracy of 68% in a similar task. This low accuracy is indicative of how convincing phishing emails can be. Indeed, when we looked at the average suspicion score given to each email (Fig. 3), there was considerable overlap between the suspicion score given to safe vs. phishing emails, with one of the real-phishing emails being among the least suspicious of all.

A unique feature of PEST is that the real-world efficacy of 84 simulated-phishing emails had been previously assessed in a field experiment in which participants were not expecting to be phished (Lin et al., 2019; Oliveira et al., 2017). Thus, by comparing real-world and lab-based efficacy, we were able to test, for the first time, whether email phishing susceptibility in the lab is related to email phishing susceptibility in the real world. In particular, we found that emails that were more effective at phishing people in the real world were also rated as less suspicious in the lab, suggesting that PEST can capture at least some of the processes underlying phishing susceptibility in real life. We note that the effect size was relatively small ($r(82) = -0.22$) and is in need of replication. At the same time, this initial finding of ecological validity lends weight to the idea of using PEST in combination with experimental neuroimaging and other physiological measures (e.g., eye tracking and galvanic skin response) in the lab, to capture the cognitive and neural processes of real-life phishing decision-making. More generally, that this particular lab-based measure of phishing susceptibility correlates with real-world

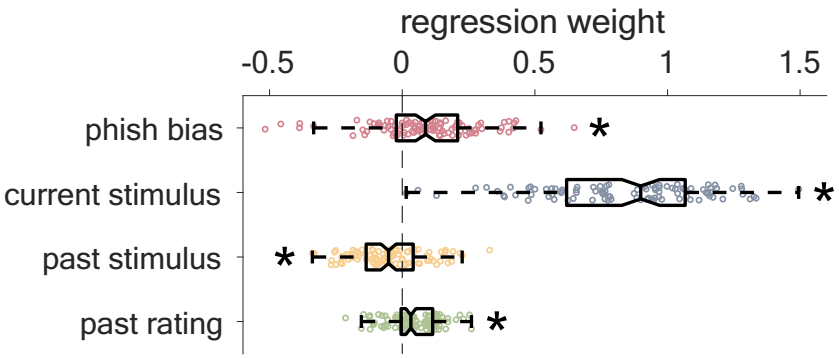


Fig. 5 Regression model of PEST behavior. Regression weights for the overall phish bias (β_{bias}), the effect of the current stimulus (β_f), the effect of the past stimulus (β_{fPast}), and the effect of the past rating (β_{cPast}) are plotted for each participant. Ratings are most strongly influenced by the average suspicion score of the current email (*blue*), but also show an overall phish bias (*red*). In addition, we see evidence of sequential effects in the form of a contrastive effect of past stimulus (yellow) and an assimilative effect of past rating (*green*). * indicates $p < 0.001$

behavior provides additional validation for other lab-based paradigms used in previous work.

Our approach is not without limitations. For example, the lab-based and real-world measures of email phishing susceptibility were acquired from different participants. This limits the scope of our conclusions such that, while we can say that PEST is effective in assessing the efficacy of individual *emails*, we cannot say that PEST is effective at determining the phishing susceptibility of individual *people*. This is important because previous research suggested that there are considerable individual differences in people's susceptibility to financial exploitation (Button, Lewis, & Tapley, 2009), with increased vulnerability to phishing associated with decreased risk perception, suspicion, memory, and positive affect, and increased conscientiousness (Halevi, et al. 2015; Vishwanath et al., 2018; Williams et al., 2017; Ebner et al., 2020). If phishing susceptibility could be accurately predicted by PEST, then a simple online test could be deployed as an assessment tool to target people in need of intervention – either to improve their ability to detect phishing emails or to provide guidance and support for their use of email (for example in older adults with cognitive decline; Boyle et al. 2019). To determine whether PEST does predict individual differences in real-world phishing susceptibility, we will need a within-subjects design to examine PEST in parallel with a direct phishing measure. These data would allow us to assess the extent to which PEST behavior captures real-world phishing behavior, and which components of PEST (e.g., regression coefficients in Eq. 1) correlate with real-world phishing decision-making across participants.

A second limitation of our approach is the relatively narrow demographics of our participants, who were undergraduate students at the University of Arizona with a mean age under 19 years old. While this group, like everyone else online, is exposed to phishing emails on a regular basis, it is not clear how their behavior is reflective of the population at large, which includes different age groups, education levels, cultural backgrounds, computer/Internet savviness, and very different settings for checking emails (e.g., in a corporate environment). Clearly expanding PEST beyond this demographic will be an important goal for future work. In part to this end, and as an invitation for others to replicate our work, we have made available all materials necessary to implement PEST (at https://github.com/zmhakim/PEST) and analyze the behavioral data (at https://github.com/bobUA/HakimEtAl_PEST_dataAndCode).

A third limitation concerns the simulated-safe emails. While these were rated as less suspicious than either the simulated-phishing or real-phishing emails, they were rated as *more* suspicious than the real-safe emails. Going forward, a key next step will be to refine these simulated-safe emails to make them appear as safe as the real-safe emails.

Finally, the original field experiment (Lin et al. 2019; Oliveira et al. 2017) also has limitations. In particular, participants consented to installing a browser plugin, which they were told was necessary to track their daily internet use (including activities such as reading the news, social media, free browsing, and checking their emails) for one hour each day. While they were not told that the true intent of the plugin was to monitor whether they engaged with phishing emails or not, nor that we would be sending them phishing emails, it is possible that the mere presence of this plugin was sufficient to affect participants' behavior. Future work should address this issue.

In addition to assessing the ecological validity of PEST, the large number of trials in our experiment allowed us to model trial-by-trial behavior with linear regression. We built a model in which participant ratings were determined by a combination of the current email, a bias towards saying an email is suspicious, and sequential effects from the last email and its rating. Of particular interest were the two sequential effects we observed: a contrastive effect for past stimulus and an assimilative effect for past rating. While sequential effects have not previously been reported in phishing tasks, they are known to occur in other sequential judgment tasks without performance feedback (Garner, 1953; Holland & Lockhead, 1968; Parducci & Marshall, 1962; Pegors, Mattar, Bryan, & Epstein, 2015; Ward & Lockhead, 1970; Wedell, Parducci, & Edward Geiselman, 1987), see (Kiyonaga, Scimeca, Bliss, & Whitney, 2017) for review.

Guided by this previous literature, the presence of these sequential effects in PEST suggests a model for real-world phishing decision-making. In particular, one way in which contrastive and assimilative sequential effects can simultaneously occur is if the judgment is made via a comparison process with recent stimuli (Holland & Lockhead, 1968). In the simplest form of this model, which considers only the stimulus and rating on the last trial, the rating on the present trial is computed by adjusting the rating from the last trial in proportion to the difference between the present stimulus and the last stimulus. Written mathematically this implies that

$$c_t = c_{t-1} + s_t + s_{t-1} \tag{3}$$

hence the past rating, c_{t-1}, has a positive, i.e. assimilative, effect on the current rating, while the past stimulus, s_{t-1}, has a negative, i.e. contrastive, effect. More generally, if the comparison process is not just with one but the average of several past stimuli, then the rating takes the form

$$c_t = s_t + \sum_i w_i (c_{t-i} - s_{t-i}) \tag{4}$$

where w_i describes the weighting given to the stimulus at time $t-i$. In line with this equation, including more terms in the regression analysis suggests that the assimilative effect of past rating and the contrastive effect of past stimuli extend multiple

trials into the past (Supplementary Fig. 3). More work remains to be done to validate such a comparison-based model of email evaluation. For example, there are many differences between the kinds of perceptual judgments used in most previous work on sequential effects (e.g., the loudness of a tone, Holland & Lockhead 1968) and the evaluation of emails, which involves combining multiple subtle perceptual and linguistic cues with complex prior knowledge about emails in general. Thus, a key goal for future research will be to better understand the integration of multidimensional cues in email evaluation, and to determine how these multidimensional cues in tandem influence sequential effects. We believe that our data provides strong initial support for this comparison model and constitutes a first step into investigation of the computational processes underlying phishing detection.

Could real-world phishing decisions be made in a similar manner? That is, are phishing decisions made by comparing the current email with exemplars of safe and phishing emails we have seen before? Such a process is certainly possible, especially in cases (such as first thing in the morning) when we work through a stack of emails sequentially in a manner quite similar to PEST. In other cases, however, (for example as emails arrive during the day) people may evaluate just one email at a time. With such sporadic email checking, there may be long delays between emails and working memory for the last email may be gone making sequential comparison more difficult. Nevertheless, people may use a comparison process even in this case, relying instead on exemplars stored in long-term memory – effectively "sampling" past experience to make the decision (Stewart et al. 2006). Detecting whether people use such comparison processes – either with recent emails or exemplars in long-term memory – will require more extensive experiments in the field as well as the lab.

Could real-world phishing decisions be made in a similar manner? That is, are phishing decisions made by comparing the current email with exemplars of safe and phishing emails we have seen before? Such a process is certainly possible, and there are many parallels between email checking, which is often done sequentially as we look at each email in turn, and the design of PEST. Nevertheless, we would need more extensive field experiments to test whether real-world email decisions are made using a comparison process, for example by sending emails in pairs such that the effect of neighboring emails in the inbox on susceptibility to a phishing email could be assessed.

Clearly, more research is necessary to advance comprehension of the complex set of processes that underlie one's decision to engage with a phishing email or not in real life. This study has taken initial steps toward the development of a tool for in-lab assessment of email phishing susceptibility, which is efficient in administration and allows data collection to inform the cognitive mechanisms of phishing detection.

Open practices statement The experiment reported in this article was not formally preregistered. De-identified data are posted on dataverse (Hakim et al. 2020). Code and data to generate all figures is posted on GitHub at https://github.com/bobUA/HakimEtAl_PEST_dataAndCode. Code to implement PEST is posted on GitHub at https://github.com/zmhakim/PEST.

Acknowledgements This work was supported by a pilot grant from the McKnight Brain Research Foundation, NSF grant SBE-1450624, and NIH grant 1R01AG057764-01A1.

Author contributions RCW, MDG, NCE, DSO, SJG, and BEL conceived the study. ZMH built the experiment with supervision from RCW and MDG. ZMH and KL designed the simulated-safe emails and collected the real-safe and real-phishing emails. NCE, DSO, and TL provided the original simulated-phishing emails and the PHIT data. ZMH ran data collection with supervision from RCW and MDG. ZMH and RCW analyzed the data. RCW wrote the paper with input from MDG, NCE, DSO, SJG, BEL, ZMH, and VL.

References

Alsharnouby, M., Alaca, F., & Chiasson, S. (2015). Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, Vol. 82, pp. 69–82. https://doi.org/10.1016/j.ijhcs.2015.05.005

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, Vol. 10, pp. 433–436. https://doi.org/10.1163/156856897x00357

Boyle, P. A., Yu, L., Schneider, J. A., Wilson, R. S., & Bennett, D. A. (2019). Scam Awareness Related to Incident Alzheimer Dementia and Mild Cognitive Impairment: A Prospective Cohort Study. *Annals of internal medicine, 170*(10), 702-709.

Button, M., Lewis, C., & Tapley, J. (2009). Fraud typologies and victims of fraud. Retrieved from https://researchportal.port.ac.uk/portal/files/1926122/NFA_report3_16.12.09.pdf on 05/16/20

Button, M., Lewis, C., & Tapley, J. (2014). Not a victimless crime: The impact of fraud on individual victims and their families. *Security Journal*, Vol. 27, pp. 36–54. https://doi.org/10.1057/sj.2012.11

Caputo, D. D., Pfleeger, S. L., Freeman, J. D., & Eric Johnson, M. (2014). Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy*, Vol. 12, pp. 28–38. https://doi.org/10.1109/msp.2013.106

Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why phishing works. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '06*. https://doi.org/10.1145/1124772.1124861

Downs, J. S., Holbrook, M., & Cranor, L. F. (2007). Behavioral response to phishing risk. *Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit on - eCrime '07*. https://doi.org/10.1145/1299015.1299019

Ebner, N. C., Ellis, D. M., Lin, T., Rocha, H. A., Yang, H., Dommaraju, S., … Oliveira, D. S. (2020). Uncovering Susceptibility Risk to Online Deception in Aging. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*. https://doi.org/10.1093/geronb/gby036

Fraud Advisory Panel. (2015). Supporting the victims of fraud: The year in review 2014–2015. Retrieved August 24, 2019, from https://www.fraudadvisorypanel.org/wp-content/uploads/2015/07/FAP-Yearly-Review-2015-WEB.pdf

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal of Experimental Psychology*, *46*(5), 373–380.

Gavett, B. E., Zhao, R., John, S. E., Bussell, C. A., Roberts, J. R., & Yue, C. (2017). Phishing suspiciousness in older and younger adults: The role of executive functioning. *PloS One*, *12*(2), e0171620.

Google. (n.d.). Disclosing vulnerabilities to protect users across platforms. Retrieved October 10, 2019, from Google Online Security Blog website: https://security.googleblog.com/2019/03/disclosing-vulnerabilities-to-protect.html

Google Safe Browsing. (n.d.). Retrieved October 10, 2019, from https://safebrowsing.google.com/

Halevi, T., Memon, N., & Nov, O. (2015). Spear-phishing in the wild: A real-world study of personality, phishing self-efficacy and vulnerability to spear-phishing attacks. *Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks (January 2, 2015)*.

Hakim, Z. M., Ebner, N. C., Oliveira, D. S., Getz, S. J., Levin, B. E., Lin, T., Lloyd, K., Lai, V. T., Grilli, M. D., and Wilson, R. C. (2020), "Evaluating the cognitive mechanisms of phishing detection with PEST, an ecologically valid lab-based measure of phishing susceptibility", https://doi.org/10.7910/DVN/DB56VY

Holland, M. K., & Lockhead, G. R. (1968). Sequential effects in absolute judgments of loudness. *Perception & Psychophysics*, Vol. 3, pp. 409–414. https://doi.org/10.3758/bf03205747

Information Security at UVA. (n.d.). Retrieved August 24, 2019, from Information Security Alerts & Warnings website: https://security.virginia.edu/security-alerts-and-warnings

Information Security University of Arizona. (n.d.). Phishing Alerts. Retrieved August 24, 2019, from https://security.arizona.edu/phishing_alerts

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.

Kelley, T., & Bertenthal, B. I. (2016). Attention and past behavior, not security knowledge, modulate users' decisions to login to insecure websites. *Information and Computer Security*, Vol. 24, pp. 164–176. https://doi.org/10.1108/ics-01-2016-0002

Kiyonaga, A., Scimeca, J. M., Bliss, D. P., & Whitney, D. (2017). Serial Dependence across Perception, Attention, and Memory. *Trends in Cognitive Sciences*, *21*(7), 493–497.

Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M. A., & Pham, T. (2009). School of phish. *Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09*. https://doi.org/10.1145/1572532.1572536

Lin, T., Capecci, D. E., Ellis, D. M., Rocha, H. A., Dommaraju, S., Oliveira, D. S., & Ebner, N. C. (2019). Susceptibility to Spear-Phishing Emails. *ACM Transactions on Computer-Human Interaction*, Vol. 26, pp. 1–28. https://doi.org/10.1145/3336141

Luo, X. Robert, Zhang, W., Burd, S., & Seazzu, A. (2013). Investigating phishing victimization with the Heuristic–Systematic Model: A theoretical framework and an exploration. *Computers & Security*, Vol. 38, pp. 28–38. https://doi.org/10.1016/j.cose.2012.12.003

Norris, G., Brookes, A., & Dowell, D. (2019). The Psychology of Internet Fraud Victimisation: a Systematic Review. *Journal of Police and Criminal Psychology*. https://doi.org/10.1007/s11896-019-09334-5

Oliveira, D., Ebner, N., Rocha, H., Yang, H., Ellis, D., Dommaraju, S., … Lin, T. (2017). Dissecting Spear Phishing Emails for Older vs. Young Adults. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. https://doi.org/10.1145/3025453.3025831

Overview of fraud and computer misuse statistics for England and Wales - Office for National Statistics. (n.d.). Retrieved October 10, 2019, from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/overviewoffraudandcomputermisusestatisticsforenglandandwales/2018-01-25

Parducci, A., & Marshall, L. M. (1962). Assimilation vs. contrast in the anchoring of perceptual judgements of weight. *Journal of Experimental Psychology*, *63*, 426–437.

Pegors, T. K., Mattar, M. G., Bryan, P. B., & Epstein, R. A. (2015). Simultaneous perceptual and response biases on sequential face attractiveness judgments. *Journal of Experimental Psychology. General*, *144*(3), 664–673.

Phishing and Scam Emails: A Realtime Database of Phishing Emails. (n.d.). Retrieved August 24, 2019, from Phishing and Scam Emails: A Realtime Database of Phishing Emails website: https://philmir.wordpress.com

PhishMe. (2016). Enterprise Phishing Susceptibility Report. Retrieved May 30, 2019, from https://cofense.com/wp-content/uploads/2017/10/PhishMe_EnterprisePhishingSusceptibilityReport_2015_Final.pdf

Rajivan, P., & Gonzalez, C. (2018). Creative Persuasion: A Study on Adversarial Behaviors and Strategies in Phishing Attacks. *Frontiers in Psychology*, Vol. 9. https://doi.org/10.3389/fpsyg.2018.00135

Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L. F., & Downs, J. (2010). Who falls for phish? *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*. https://doi.org/10.1145/1753326.1753383

Smart People Easier. (2014). Smart people easier to scam. Retrieved August 24, 2019, from http://www.ultrascan-agi.com/public_html/html/pdf_files/Pre-Release-419_Advance_Fee_Fraud_Statistics_2013-July-10-2014-NOT-FINAL-1.pdf

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive psychology*, 53(1), 1–26.

Test and Optimize your Emails for the Inbox. (n.d.). Retrieved August 24, 2019, from https://glockapps.com/spam-testing/

UCLA Information Security Office. (n.d.). PHISH BOWL/PHISHING SCAMS. Retrieved August 24, 2019, from https://www.it.ucla.edu/security/alerts/phishing-scams

Vishwanath, A. (2015). Examining the Distinct Antecedents of E-Mail Habits and its Influence on the Outcomes of a Phishing Attack. *Journal of Computer-Mediated Communication*, Vol. 20, pp. 570–584. https://doi.org/10.1111/jcc4.12126

Vishwanath, A., Harrison, B., & Ng, Y. J. (2018). Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research, 45*(8), 1146–1166.

Ward, L. M., & Lockhead, G. R. (1970). Sequential effects and memory in category judgments. *Journal of Experimental Psychology*, Vol. 84, pp. 27–34. https://doi.org/10.1037/h0028949

Wedell, D. H., Parducci, A., & Edward Geiselman, R. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology*, Vol. 23, pp. 230–249. https://doi.org/10.1016/0022-1031(87)90034-5

Williams, E. J., Beardmore, A., & Joinson, A. N. (2017). Individual differences in susceptibility to online influence: A theoretical review. *Computers in Human Behavior, 72*, 412–421.

Williams, E. J., Hinds, J., & Joinson, A. N. (2018). Exploring susceptibility to phishing in the workplace. *International Journal of Human-Computer Studies*, Vol. 120, pp. 1–13. https://doi.org/10.1016/j.ijhcs.2018.06.004

Wilson, R. C. (2018). *Sequential choice effects predict prevalence-induced concept change*. https://doi.org/10.31234/osf.io/75bpy

Wood, S., Liu, P.-J., Hanoch, Y., Xi, P. M., & Klapatch, L. (2018). Call to claim your prize: Perceived benefits and risk drive intention to comply in a mass marketing scam. *Journal of Experimental Psychology: Applied*, Vol. 24, pp. 196–206. https://doi.org/10.1037/xap0000167

Yan, Z., & Gozu, H. Y. (2012). Online Decision-Making in Receiving Spam Emails Among College Students. *International Journal of Cyber Behavior, Psychology and Learning*, Vol. 2, pp. 1–12. https://doi.org/10.4018/ijcbpl.2012010101