Summer 2023

An Exploration of the Low Prevalence Effect During Phising Detection

Sherry J. Wei
San Jose State University

AN EXPLORATION OF THE LOW PREVALENCE EFFECT DURING PHISHING DETECTION

A Thesis

Presented to

The Faculty of the Department of Psychology

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

Sherry J. Wei

August 2023

The Designated Thesis Committee Approves the Thesis Titled

AN EXPLORATION OF THE LOW PREVALENCE EFFECT DURING PHISHING DETECTION

by

Sherry J. Wei

APPROVED FOR THE DEPARTMENT OF PSYCHOLOGY

SAN JOSÉ STATE UNIVERSITY

August 2023

Evan Palmer, Ph.D.	Department of Psychology
Sean Laraway, Ph.D.	Department of Psychology
Christina Tzeng, Ph.D.	Department of Psychology

ABSTRACT

AN EXPLORATION OF THE LOW PREVALENCE EFFECT DURING PHISHING DETECTION

by Sherry J. Wei

Phishing attacks are attempts to obtain individual credentials or other private information through deception, usually in email format. As the Internet becomes increasingly intertwined with everyday lives, such attacks are on the rise, threatening individuals and businesses alike. Existing anti-phishing training measures fail to address possible prevalence effects on detection performance: in tasks where targets appear rarely, participants have heightened miss rates. This low prevalence effect could be present in phishing detection because phishing emails are observed much less frequently than legitimate emails. Emerging research has reported observing heightened miss rates as a function of phishing email rarity. This study aimed to replicate those findings with improvements to both the internal and external validity of the task design by using real-life emails as stimuli and increasing the stimulus set size. Participants attempted to identify phishing emails among normal emails and were randomly assigned to one of four phishing prevalence conditions: 1%, 3%, 5%, and 20%. Sensitivity did not significantly differ between prevalence groups, nor did we observe significant differences in criterion or miss rates. Limitations of the study include not accounting for English fluency, which is a possible covariate. More research is needed to understand whether the low prevalence effect is observed during phishing detection.

DEDICATION

To my sister, parents, and brothers. You give me the momentum to succeed.

ACKNOWLEDGEMENTS

I extend my warmest gratitude to my advisor, Dr. Evan Palmer. Your guidance, enthusiasm, and insight were invaluable to both the culmination of this work and to my personal growth. I always enjoyed the friendly notes you slipped between your well-worded and thoughtful advice. Thank you for always supporting a healthy positive mindset!

To my sister, Sally: it is no exaggeration to say that this thesis would not have been possible without your ingenious scripts. Thank you for the late nights and long hours you spent creating a scalable solution.

To Dr. Laraway and Dr. Tzeng: thank you both for sharing your expertise, which helped shape decisions along the way.

To my cohort—Stephanie, Alex, Tyler, Sol (and the list goes on!): thank you for the board games, the study sessions, the jokes and emotional support…. It would have been a very different experience without you all on similar journeys beside me.

To my brothers and parents: thank you for your unwavering support, random check ins, and all the emails. You raise me up!

Finally, to everyone who took part in feedback sessions, the pilots, and the full-length study: thank you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Nearly $3.5 billion was lost to internet crime in the United States in 2019 alone (Federal Bureau of Investigation, 2020). Of all the attack methods, phishing was one of the most frequently reported. Phishing is a social engineering tactic whereby the attacker poses as a legitimate entity using an electronic medium, with the aim of stealing an individual's credentials or other private information (Jagatic et al., 2007). Financially driven phishing attacks are on the rise, targeting both companies and individuals. In 2022, the Federal Bureau of Investigation reported that phishing-type attacks were the top crime type that they received complaints about, comprising of 22% of all crime reports received by the Internet Crime Complaint Center that year and causing $44 million in losses. In addition, Verizon reported in its annual Data Breach Investigations Report in 2020 that one-fifth of investigated data breaches involved phishing.

Phishing attacks include messages that attempt to incite the user to click on a provided link or attachment. The content and structure of these messages vary widely, as do the appearance of the linked false websites. Phishing attacks occur through several channels, appearing as emails, texts, social media posts, and phone calls. Despite having high variability, phishing attacks often share common characteristics. These characteristics, called *cues*, include the sender's address, vague language, and invocations of urgency or threats (Downs et al., 2006). Any number of these cues may be present or absent in a phishing attack.

Phishing attacks are advantageous to the attacker because thousands of phishing attempts can be sent out with ease, and just one successful phish may be enough to lead to a large data

breach. Given the polymorphous nature of this ongoing threat and the critical role of human defenders, there has been a growing body of research aimed at better understanding and combating phishing from a perspective centering around human behavior in relation to detecting phishing attacks. A large body of research has focused on phishing emails over other mediums since that is the most common attack vector (Accenture, 2019).

A Human-Centered Approach

The role of human decision making in vulnerability to phishing attacks has prompted researchers to explore human variables related to falling victim to a phishing attack. Initial research on phishing susceptibility has studied decision making strategies, familiarity with technology or phishing, and individual differences to better understand phishing susceptibility (see Moreno-Fernández et al., 2017 and Williams et al., 2017 for a review). Downs et al. (2006) examined the decision-making strategies of individuals with low computer security knowledge in relation to phishing detection. They sought to understand what thought processes or criteria individuals use during phishing email classification, and used a method of measuring phishing susceptibility that has been widely adopted in the field. The susceptibility measure utilizes a yes-no task with a stimulus set of phishing or legitimate emails, where participants are asked to categorize emails as one category or the other. Participants were additionally given a role to play in order to contextualize the emails (a scenario-based task). The stimulus set they created included five phishing and three legitimate emails, with relevant details such as sender address and link or URL text noted down. During the evaluation of each email, participants were asked to talk aloud about reasons for their actions. Following completion of the phishing detection task, participants

were interviewed about their security awareness and asked to identify the characteristics that dictated whether they should trust the email's content.

Downs et al. (2006) found that awareness of security cues like indicators for website security levels did not affect the level of caution in approaching the emails. Often, participants would misinterpret security cues, leading to faulty evaluations. Other studies (Alsharnouby et al., 2015; Egelman et al., 2008; Kelley & Bertenthal, 2016) supported these findings, observing that participants spent very little time looking at security warnings and tended to ignore them.

Kelley and Bertenthal (2016) examined human interaction with security warnings on browsers in the second stage of a phishing attack, where users face a website that prompts them for user information while masquerading as a legitimate site. Participants were instructed to decide between logging in or backing out of the website based on how secure they believed the website to be. They were incentivized by a monetary reward to act as quickly and accurately as possible. A tendency to log in regardless of security warnings was observed. Those with high security knowledge, who logged in less frequently if there were security warnings present, were not better at determining the legitimacy of websites in the absence of security warnings. Kelley and Bertenthal suggested that in everyday use, users were accustomed to ignoring security warnings. This is perhaps because security warnings are often present on benign but insecure websites, and only rarely indicate a malicious one. Additionally, insecure hosting is no longer a reliable indicator of a malicious website: 68% of phishing websites reported internationally in the third quarter of 2019 were hosted using secure HTTPS protocol (Anti-Phishing Working Group, 2019). This highlights an issue

where individuals become vulnerable if an expectation regarding characteristics of a phishing attempt is violated.

Security awareness, while important, is not enough to prevent vulnerability to phishing. In an exercise conducted at West Point Academy, cadets who received four hours of security awareness training were deceived by benign phishing emails mere hours following the training sessions (Ferguson, 2005). High email load and habitual email use were also linked to a higher likelihood to be phished (Vishwanath et al., 2011). To simulate the circumstances of a real phishing attack, embedded phishing exercises were appended to phishing awareness programs (e.g., Kumaraguru et al., 2007). Embedded phishing exercises involve sending unannounced benign phishing emails with links to educational material, should individuals click on the link. These test emails are sent every few weeks or months (Siadati et al., 2017), reflecting the rarity of real phishing email occurrences. This training strategy is commonly employed in companies and organizations today.

General susceptibility to phishing attacks is reported to be decreasing, from 14.1% in 2015 and 12.6% in 2016 (Cofense Inc., 2019) down to 10.8% in 2017 (Cofense Inc., 2017) and 9.8% in 2019 (Cofense Inc., 2019). However, each year a smaller decrease is observed. Currently employed strategies are seeing a gradual plateau in beneficial results. A phenomenon known as the low prevalence effect may be relevant for understanding this trend, given that embedded training does not address phishing attack rarity in a manner that encourages individuals to guard against it. The next sections will describe the literature on the low prevalence effect.

The Low Prevalence Effect

The low prevalence effect (LPE) describes a situation where individuals tend to miss targets disproportionately often when those targets are rare compared to when they are relatively frequent (Wolfe et al., 2005). For example, while visual search tasks in the laboratory often include targets at a 50% prevalence rate, in real world visual searches such as during airport baggage screening, targets appear much less often. Wolfe et al. (2005) designed a baggage screening task varying the number of objects viewed in each trial and the prevalence rate of targets to study how visual search behavior interacted with target rarity. Participants were assigned to conditions with 1%, 10%, or 50% target prevalence. In the 1% condition, participants viewed over 2,000 trials, but only saw 20 targets. In the 10% and 50% conditions, participants viewed over 200 trials with an unspecified number of targets. Results revealed that error rates increased as prevalence decreased. For example, when target prevalence was 50%, participants missed targets 7% of the time, but when target prevalence was 1%, participants missed targets 30% of the time. Response time data revealed that participants were inclined to quickly abandon their search when the target was rare, with faster reaction times for target-absent responses than for target-present ones.

Subsequent studies suggested that this behavior and decline in performance is not due to motor errors from fatigue during difficult search tasks. During easy low prevalence tasks, participants tasked with indicating target-absent or target-present by pressing buttons may develop a somewhat automatic response for indicating target-absent (Fleck & Mitroff, 2007). These motor response errors can be corrected by allowing participants to double-check their work and change their answer (Fleck & Mitroff, 2007). However, during difficult search

tasks, the LPE persists even if responses are correctable (Van Wert et al., 2009; Wolfe et al., 2007). Additionally, when participants have difficulty detecting a target's presence through perception, they tend to choose a response based on the perceived prevalence of the target (Schwark et al., 2012). The LPE is also not due to the target being too difficult to detect. Hout et al. (2015) found that participants often missed rare targets even when they looked at them directly, as measured by eye tracking.

The Low Prevalence Effect with Trained Observers

Alarmingly, even highly trained, professional observers display the LPE (Evans et al., 2013; Sawyer & Hancock, 2018; Wolfe et al., 2013). Evans et al. (2013) examined prevalence effects during mammography screening by radiologists in a clinical setting. Expert mammographers were shown 50 positive and 50 negative cases within their normal workflow over the course of nine months, leaving the usual (low) prevalence rate unaffected. These cases were later reviewed in a single session by six radiologists who had also participated in the low prevalence clinical condition. Miss rates were significantly higher in the low prevalence clinical condition than in the high prevalence condition, suggesting a prevalence effect even among trained individuals.

The LPE has not only been found with trained observers in static search tasks but also in dynamic tasks. Beanland et al. (2014) examined whether prevalence effects would occur during a driving simulation. Participants with driving experience were instructed to look for buses and motorcycles while watching vehicles approach and pass by. In one condition, motorcycles were rare and buses were highly prevalent, while in the other, the opposite was true. While participants rarely missed targets, they were significantly faster at detecting high

prevalence targets compared to low prevalence targets regardless of target size, thus displaying a prevalence effect.

Given that the low prevalence effect has been observed in multiple contexts and persists with training, researchers have proposed explanations for the LPE in the interest of aiding mitigation attempts. Several proposed explanations of the LPE utilize signal detection theory (SDT), which provides a framework for understanding how people weigh information and make decisions when detecting targets among ambiguous stimuli (Macmillan & Creelman, 2004; Martin et al., 2018). The following sections provide an explanation of SDT's core concepts and tie those concepts to the low prevalence effect.

Signal Detection Theory

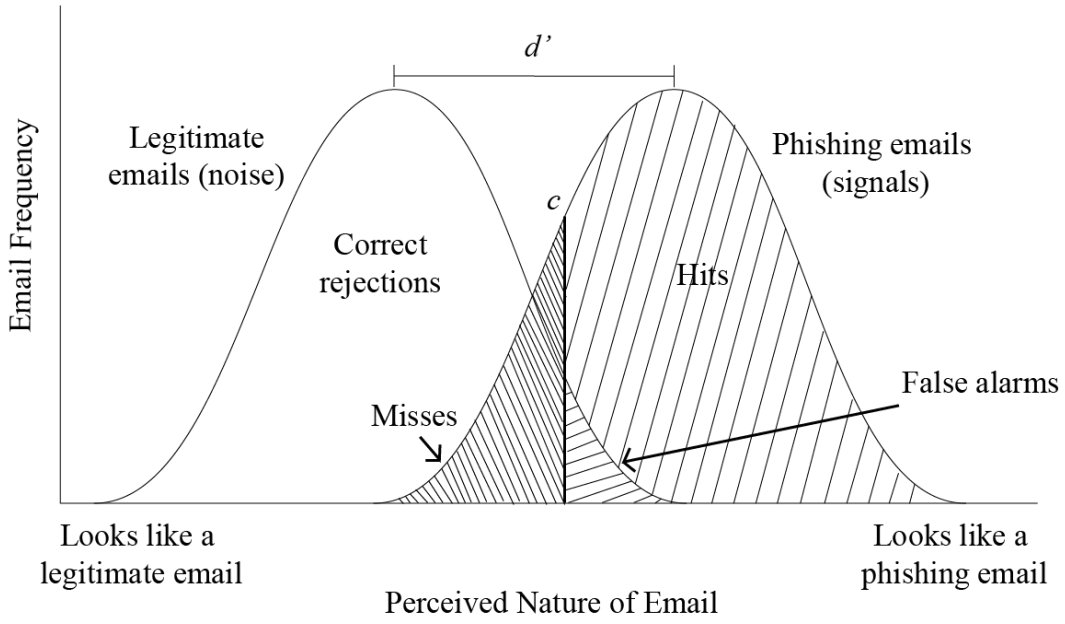
SDT divides stimuli in the environment into targets, called signals, and non-targets, called noise. To account for ambiguity and variability in stimulus strength, signal and noise stimuli are modeled as two normal distributions that partially overlap. Figure 1 provides a visual representation of this.

The signal distribution has a higher mean than does the noise distribution. An individual's performance is affected both by their ability to recognize the signal (sensitivity) as well as how likely they are to answer one way or another when given an ambiguous stimulus (criterion; *c*). Sensitivity (*d'*) is represented by the distance between the peaks of the noise and signal distributions. An unbiased criterion (*c*) is the value at the point where the distributions overlap, where stimuli with strengths falling to the right of the criterion will elicit signal-present responses, while stimuli with strengths falling to the left of it will elicit signal-absent responses. Shifting the criterion toward the left or the right signifies a bias

toward responding with more signal-present or more signal-absent decisions, respectively. These concepts can be measured by observer response data.

Figure 1

Signal and Noise Distributions in Signal Detection Theory



Note. d' represents sensitivity, and *c* represents the criterion.

Correctly identifying a signal is a hit, while misidentifying a signal as noise is a miss. Correctly identifying noise is a correct rejection, while misidentifying noise as a signal is a false alarm. Sensitivity is determined by how accurately the observer makes signal-present responses; accurate performance comes from having both a high hit rate and a low false alarm rate. This model allows researchers to quantify performance and decision making during visual search tasks. The next section describes explanations for the low prevalence effect using SDT.

Explanations for the Low Prevalence Effect

One possible explanation for the low prevalence effect is that perhaps individuals miss targets more often due to responding more quickly, a concept known as a speed-accuracy tradeoff (see Heitz, 2014 for a review). Another explanation is that perhaps individuals have lower sensitivity because of the rarity of the target, while a third explanation is that low prevalence leads to a shift in criterion. Wolfe and Van Wert (2010) conducted a set of experiments to examine these possibilities. To examine whether a speed-accuracy tradeoff had occurred, participants searched for targets that were either present in 98% of the trials or in 50% of them. Performance for errors during target-absent trials was examined. In the 98% prevalence condition, an elevated false alarm rate was observed compared to in the 50% prevalence condition. However, rather than observing faster reaction times in the high prevalence condition as a speed-accuracy tradeoff would suggest, no such relationship was observed. Instead, in the 98% condition where absent targets were rare, target-absent responses were greatly slowed compared to in the 50% condition. Thus, Wolfe and Van Wert concluded it was unlikely that a speed-accuracy tradeoff would explain the low prevalence effect.

Reported in the same paper, a follow-up experiment examined the impact of varying target prevalence on sensitivity and criterion. Could lowered sensitivity be an explanation for the low prevalence effect? Participants searched for targets in 1000 trials where the prevalence of the target varied from high to low and back to high over the course of the study. Results showed that criterion changed systematically with prevalence rate, while sensitivity did not. As prevalence decreased, criterion values were observed to increase,

meaning that participants gave relatively more target-absent responses when they realized target prevalence was decreasing. Additionally, only target-absent response times were observed to change alongside prevalence. In other words, during the trials, participants would adjust their understanding of the prevalence and modify their criterion as well as how quickly they would decide on target absences. These results suggest that the low prevalence effect can be explained as a shift in criterion, where signal rarity predisposes individuals to identify ambiguous stimuli as noise, thus leading to a heightened miss rate. Should the low prevalence effect be observed during phishing detection tasks, this understanding of the LPE could inform improvements to training program implementation or design.

Phishing and Low Prevalence

Research on phishing detection typically either uses a 50% prevalence rate (e.g., Downs et al., 2006) or a real-world prevalence rate of about 1% (Williams et al., 2017). Studies exploring the low prevalence effect in the context of phishing detection are few and lack consensus. Aiming to examine the impact of prevalence on phishing detection performance, Sawyer and Hancock (2018) designed an inbox simulation exercise where participants were tasked with responding appropriately to received emails. Assuming the role of an employee at a dummy company, participants received emails requesting that they send back a document or accept an attached file. Participants were allowed to take as much time as they needed and were instructed to respond appropriately. To respond, participants selected one of three options: report as suspicious, attach a document, or file the attached PDF.

Thirty participants were randomly assigned to one of three conditions, where the percentage of phishing emails shown was 1%, 5%, or 20%. In each condition, 300 emails

were shown sequentially and in full. Phishing and legitimate emails differed with a consistent cue in the form of sender domain suffix (".tv" versus ".com"), while emails containing malicious attachments additionally consistently ended in ".exe" instead of ".pdf". In the 1% condition, results revealed that although participants took more time to complete the task, their response accuracy was lower compared to the other two conditions, where participants took less time while having higher accuracy rates. This suggests a prevalence effect at the 1% prevalence rate, but not at the 5% or 20% rates. Sawyer and Hancock (2018) thus suggested that, as automated phishing detection technology improves, human phishing detection as the final line of defense would falter due to this phenomenon.

Despite these potent findings, the study has several limitations in both task and stimulus design. In the study, phishing emails consistently included an obvious cue, but in real world settings, phishing emails do not have a consistent set of cues and may even take advantage of "safe" cues like valid sender addresses to deceive their victims. Additionally, during real world situations where individuals are tasked with sifting through a large volume of emails, they do not view each email sequentially but rather many at once through a preview pane. Of further note is the low number of trials utilized: with 300 trials at a 1% prevalence rate, only 3 data points were obtained per participant.

A more recent study by Sarno and Neider (2022) showed 72 participants 100 emails at 5%, 25%, or 50% prevalence, and found that, contrary to Sawyer and Hancock's findings, sensitivity decreased with prevalence. They did not find the low prevalence effect at 5% prevalence. While their stimulus set better approximated real-life emails, their low set size limits the results of their study.

Study Objectives

The present study aims to improve on the above points by employing a novel task design as well as a larger and more varied stimulus set. Instead of showing emails sequentially, multiple emails will be shown at once in an approximation of an inbox preview pane. This method not only is more ecologically valid but also allows researchers to display more trials at once. However, because people are scanning an inbox preview pane with dozens of emails displayed at once, it is not possible to obtain response times per trial. Because heightened miss rates are the primary indicator of the low prevalence effect, this drawback was judged to be minor in comparison to the benefits.

A threshold for the low prevalence effect in the context of phishing is also not clear. Sawyer and Hancock (2018) found the low prevalence effect in phishing detection at a phishing prevalence rate of 1%, but not at 5%. However, in other research, the low prevalence effect has been observed with 2% prevalence (e.g., Wolfe & Van Wert, 2010) as well as at 5% (e.g., Hout et al., 2015). Methodologically speaking, the number of trials needed to ensure adequate power differs greatly depending on the target prevalence, with 5% allowing for substantially fewer experimental trials, overall. While obtaining an exact threshold is beyond the scope of this study, the present study aims to examine the impact of low prevalence on performance in phishing detection tasks.

To summarize, the current study has two objectives:

1. Investigate the relationship between target rarity and phishing detection using a more varied stimulus pool and an inbox simulation task that shows multiple stimuli at once.

2. Explore at what prevalence rates the low prevalence effect manifests, in the context of phishing.

Hypotheses are as follows:

H1: Sensitivity (*d'*) will not vary across prevalence conditions.

H2: A positive criterion shift will occur as prevalence decreases (in the 1%, 3%, and 5% prevalence conditions). Participants will display heightened miss rates while maintaining false alarm rates similar to those from participants in the high prevalence 20% condition.

H3: A main effect of prevalence will be observed on miss rates (miss rates will be higher in lower prevalence conditions).

Findings may suggest improvements to existing phishing training systems, such as embedded training. This study additionally includes the creation of a database of phishing and legitimate emails that may be useful in future research on phishing.

Methods

Participants

Participants were undergraduate students at San José State University (SJSU), recruited through SONA Systems, who were compensated with course credit. SJSU's Internal Review Board (IRB) reviewed and approved this study's planned methodology and analyses. Participants gave their informed consent prior to beginning the study.

Participants were screened for fluency in English, which was measured by one item like an attention check ("Please select the answer that is third from the top"). While miscategorization is a concern for this method, responses were not expected to differ egregiously from true fluency levels (for example, where an individual cannot understand the meaning of simple sentences but responds that they are fluent). Additionally, individuals who cannot quickly respond correctly to the question may apply excessive amounts of time to complete the thousand-trial task and would not be able to receive the appropriate level of compensation in credit, which was capped at 1.5 hours.

A statistical a-priori power analysis was performed for sample size estimation using GPower 3.1, based on data from Sawyer and Hancock (2018) ($N = 30$), comparing accuracy and response time on a composite of phishing email prevalence levels. Partial eta-squared in this study was .27, converting to an effect size of .61. With an alpha = .05, power = .80, and a more conservative but still large effect size of .40, the projected sample size needed with this effect size is approximately $N = 76$ for a between-subjects comparison with four groups.

Reponses that were abandoned before survey submission or responses where participants spent an unrealistically low amount of time on the task while performing poorly were

excluded from analysis. Exclusion criteria are discussed further in the results section. The final sample size was 82, with 21 responses each in the 1%, 5%, and 20% conditions and 19 in the 3% condition. Demographic information such as age and gender were not collected as no significant effect on phishing detection was expected after a review of the literature: research on the effects of age and gender on phishing detection appears inconclusive and small effect sizes were reported (Abbasi et al., 2016; Canfield et al., 2016; Chen et al., 2018; Kleitman et al., 2018; Sheng et al., 2010).

Materials

A commonly used paradigm in phishing detection studies involves showing participants emails in their entirety. However, when sifting through large volumes of unread emails in everyday situations, people often use the inbox preview pane to filter emails. To approximate this situation more closely, images with ten emails each were created and displayed through Qualtrics.com, an online survey-making site. Because having a one-line preview would not provide enough information on the email body, email content previews were expanded to a maximum of five lines to allow links to appear. Additionally, because Qualtrics does not allow hover-text, where sender addresses would typically be displayed, sender addresses were placed below sender names. Participants were shown five 10-email images per page, with each email mapped to a hotspot that could be selected or deselected. Figure 2 displays an example image.

Figure 2

Example Email Stimuli and Format Shown to Participants

The Onion Newsletter <newsletter@email.theonion.com>	**Biggest Hidden Costs of Giving Birth in America!** Breaking News! Biggest Hidden Costs of Giving Birth in America! Breaking News! Bored Census Bureau Employee Changes Every Ohio Resident's Name to Laura.
Magazines Direct <info@email.magazinesdirect.com>	**Top Tech, Low Prices!** Save up to 61%, subscribe now! Hi subscriber, keeping up with the rapid pace of the techology world has never been easier! We are now offering our best price on a range of titles. Subscribe today!
Inside CNN <insidecnn@mail.cnn.com>	**Meet the Team Behind CNN's New Year's Eve Live!** An exclusive, inside look with your free CNN account! Who's behind the cameras at CNN's New Year's Eve live? Whether it's a big party or a quiet night at home, New Year's Eve is often celebrated with the TV turned to CNN. Read the full article now!
Clark's Shoes <clarks@email.clarksusa.com	**HURRY: Last Weekend of Final Clearance** Shop and save up to 60% OFF sale styles. Shop Women's | Shop Men's FREE SHIPPING & FREE RETURNS Winter Clearance up to 60% off Shop Womens Shop Mens Shop Kids
Customer Support <customer@support-amazon.com>	**Re: [Report Confirm Amazon Info]: Your account has been locked and holds all your last orders. Thursday, March 9, 2023 EST** This is your last warning. Confirm your Amazon account information or your account will be deleted. Click here to log in and prevent your account from being locked. Thank you.
Michael J. Isip <member@email.kqed.org>	**Last Chance for a Tax Deductible Donation!** There is still time---make your 2022 donation now! Hello, as I reflect on the past year, I am so grateful to the support form our KQED community. It is because of our viewers and listeners like you that we are able to share this year's top stores in Northern California and beyond . I am asking you to consider making your first tax-deductible donation to KQED in the next few hours. It will be matched dollar-for-dollar. Your support means everything to us. Sincerely, Michael J. Isip President & CEO P.S. The deadline to make a tax-deductible gift is midnight today! Donate now!
Half Price Books <mail@hpbdirect.com>	**Required Reading:** You Need These New Arrivals. New arrivals at your local HPB, stock varies by store. I Have Some Questions For You by Rebecca Makkai "Unput-down-able and unforgetting, Makkai has written the book of the season - Andrew Shawn Greer, author of Less and Less is Lost. Check it out
TurboTax <TurboTax@em1.Turbotax.intuit.com>	**(Sign in) Your refund is a few clicks away.** Your max refund is waiting. You're only a few steps away from filing. Get started You have a head start! Act now! As a returning user, we securely transfer your info from last year so you're already one step closer to your refund.
sam's club <asdp82@saamsclub.com>	**Sams Club reward for you-Open immediately!** Congratulations to user sam's clube! Take this short, 30-second sam's club! survey to select one of our exclusive reward offers notice : this offer is only available for this email! Respond now for rewards !
Swarovski <swarovski@newletter.swarovski.com>	**Discover Millenia icons** Sophistication and simplicity are in perfect harmony in every single piece of jewelry perfection that is our magical Millenia range. Discover icons

Note. An image of emails shown to participants as part of their training on the task. Note the 5th and 9th emails from the top, purported to be from Amazon customer service and Sam's Club respectively. Both contain multiple characteristics common to phishing attempts, such as an urgent call to action, a vague or implausible premise, and wrong sender addresses.

A total of 1084 real-world emails were collected for use in this study. Of those, 52 were phishing emails, and the remaining 1032 were legitimate. Of the legitimate emails, 980 were filler and 52 were paired with phishing emails.

Real-world phishing emails were obtained from Cornell University's Phish Bowl (https://it.cornell.edu/phish-bowl), a database of reported and confirmed phishing emails sent

to individuals on the university campus, as well as the researchers' own email inboxes. Phishing emails had to include a link in the body text of the email as well as arrive from a fake sender address. Sender usernames for these emails were random combinations of letters and numbers to make them easier to identify. In addition, domain addresses utilized common tactics such as number and character substitutions.

Unfamiliarity with what a legitimate version of a phishing email might look like is a possible reason for missing a phish, so to help mitigate this, each phishing email was paired based on content to a legitimate email. Both the sender and the content were matched as closely as possible. In cases where the sender would not legitimately include certain content, the same content from a different sender or from the same sender saying they would not collect personal information was substituted. Finally, additional legitimate emails were collected to achieve the low prevalence rates required for the design.

Real-world legitimate emails were obtained from a variety of subscription services (e.g., Spotify, Netflix) and from the researchers and their associates' email inboxes. Personal greetings, sent-to addresses, and other personal information present in emails were removed (or replaced with fake information) from all emails to preserve privacy and so that the use of a scenario could be avoided. Studies on use of scenarios in experiments suggest a lack of generalizability to real world situations for certain results (e.g., Kim & Jang, 2014).

Emails were assigned numbers and ordered into sets of 10 based on numbers drawn from a random number generator. Phishing emails were uniformly distributed across all images in a condition such that in the low prevalence conditions, each image would have either zero or one phish out of ten, while in the 20% condition, each image would have two. Additionally,

phishing emails' placements in the images were distributed uniformly, such that each row across the condition had the same number of phishing emails. Legitimate emails that were matched with phishing emails were always present in the same image. Image order was further randomized using Qualtrics' within-block question randomization function to avoid possible order effects, such as participants always seeing a phish in the beginning of one task in one low prevalence condition, but not others. All phishing emails present in lower prevalence conditions were also present in higher-prevalence conditions.

Pilot Studies

Two pilot studies were conducted. The first aimed to test the mechanics of proposed stimulus presentation strategy and assess the difficulty of the detection task, while the second aimed to determine expected completion times of the study, assess the difficulty of the email stimuli after modifications, and further refine mechanical aspects of the study.

The first study iterated on an earlier design of stimulus presentation, which was to show 50 emails per image and fewer lines of text per email in Qualtrics. It also served to inform decisions on the number of stimuli to show per condition, the direction of stimulus creation, and improvement on the scalability of the design. Six participants, sourced by word of mouth through the researchers' networks, were shown a subset of stimuli depending on the condition they were in. The low prevalence conditions were reduced to the 1% and 5% conditions, and each included 250 out of the planned 1000 emails. The high prevalence condition, 20% prevalence, included 50 emails total out of the planned 250 emails. Images were created in Adobe InDesign, a desktop publishing and typesetting software application

produced by Adobe Systems (https://adobe.com/products/indesign). One block of questions was allotted for each condition.

Following this study, images were reduced from 50 emails to 10 emails per image to preserve a minimum level of readability. The image creation process also moved away from manual creation to automatic creation by a script coded in Python 3.7.1 (https://www.python.org/about/) to improve scalability and iteration speed. An informal survey of hit and miss rates for the control condition showed an unexpectedly high miss rate across conditions, suggesting that the differences between signal and noise needed to be made clearer. Thus, for the next pilot study, phishing emails were modified to have random strings of characters and letters as sender usernames that, if noticed, were obviously not legitimate.

The second pilot study included a training block prototype, 4 condition blocks, and a post-task block. Both the stimulus set's images and the survey were constructed using Python scripts for improved scalability. Fifteen participants, sourced through word of mouth, completed the survey. An informal analysis of sensitivity, criterion, and miss rates suggested that the detection task was reasonably challenging and neither too easy nor too difficult. The survey construction method and training block design were further refined prior to the full-length study.

Design

This between-subjects study examined three dependent variables derived from participant response data—sensitivity (*d'*), criterion (*c*), and miss rates—and one independent variable, phishing prevalence measured in percentage of total displayed emails, on four levels: 1%,

3%, 5%, and 20%. The 1%, 3%, and 5% conditions were considered low prevalence, while the 20% condition was considered high prevalence. Participants were randomly assigned to one of four conditions through Qualtrics' built-in random assignment feature.

Procedure

All participants were asked to complete the task in a single online session. Participants first answered a 2-item screener assessing reading comprehension and ensuring that they accessed the survey from a desktop or laptop browser; then, they read a consent notice, to which they indicated consent or non-consent. After this, participants received training on both phishing email characteristics and the task. Training materials can be found in the appendix. In the training section, they first received instructions on common phishing cues. Then, they were shown an image of 10 emails, 2 of which were phishing messages to match the high prevalence condition and were instructed to click on the emails that appear to be phish. They were given feedback on their responses in the form of red boxes indicating which emails in the image were phishing, accompanied by a description of what characteristics were present in each and a warning that phishing emails vary in how many characteristics they include.

Following the training block, participants were randomly assigned to one of four groups, varying in level of phishing email prevalence: 1%, 3%, 5%, and 20%. In all conditions, participants were instructed to clean out their inbox and select any emails that appeared to be phishing emails. They were allowed to self-pace and take breaks in the event of eye strain or other reasons. Those in the 1%, 3%, or 5% condition were shown 1000 emails split into 100 images with ten emails each, with five images per page. Of that thousand, just 10, 30, or 50

respectively were phishing emails. Participants in the 20% condition were shown 250 emails in blocks of 5 ten-email images per page, of which 50 were phishing emails. After completing the final trial, participants reported when they last were trained on phishing characteristics, if ever, and how frequently they received such training. Participants in the longer, low prevalence conditions (1%, 3%, and 5%) took on 83.72 minutes, 84.19 minutes, and 81.54 minutes on average respectively, to complete the task after two outliers each were removed from the 1% and 3% conditions. Participants in the shorter, high prevalence condition took 36.30 minutes on average.

Analysis Methods

Hit rates (proportion of phish detected) were calculated for all emails per participant. Miss rates were calculated for phishing emails (proportion of phish that went undetected), while false alarm rates were calculated for legitimate emails that were paired with those phishing emails (proportion of paired legitimate emails that were incorrectly identified as phish) as well as for legitimate emails overall. Standardized scores of these rates were used to 'obtain participant sensitivity and criterion values through the following formulas: $d' = Z(H) - Z(F)$; $c = -[Z(H) + Z(F)]/2$, where H represents hit rate, F represents false alarm rate, and Z represents the area under the cumulative normal distribution corresponding to each proportion. In cases where there were too few false alarms to conduct an analysis, only miss rates were analyzed. Time spent on each page and overall performance on the task were used to determine cases where individuals responded more quickly than possible.

For hypothesis 1, which expects no difference in sensitivity between conditions, an alternative to traditional hypothesis testing must be used. This alternative is known as a

Bayesian approach. Bayesian analyses aim to quantify and weigh evidence supporting competing hypotheses against each other through the computation of Bayes factor (Ly et al., 2020). The Bayes factor (K) is a ratio of the likelihood of an alternative hypothesis over the likelihood of the null hypothesis. In general, interpretations of Bayes factor follow a general pattern: data reveals substantial evidence supporting the alternative when K is much greater than 1, strong evidence supporting the null when K is close to 0, and no particular evidence supporting either hypothesis when K is close to 1 (Dienes, 2014). An inverted Bayes factor ($1/K$) value was obtained to compare the likelihood of no difference in sensitivity to the likelihood of a difference, where likelihood of a difference is considered the alternative hypothesis. Using Kass and Raftery's (1995) criteria, values of $1/K \geq 3.2$ were considered substantial evidence supporting the null hypothesis model while values of $1/K \leq 3.2$ were considered substantial evidence supporting the alternative hypothesis model.

A one-way between-subjects ANOVA was conducted on the rates and measures calculated from the collected data to test hypotheses 2 and 3, which state that criterions are expected to be higher in lower prevalence conditions and that miss rates will change in relation to prevalence. Additionally, for hypothesis 2, comparisons between the 1% and 20%, 3% and 20%, and 5% and 20% were planned.

All analyses were conducted using JASP (https://jasp-stats.org/). JASP is a statistics software supported by the University of Amsterdam that makes both traditional frequentist statistical analyses and computationally complex Bayesian analyses more accessible. Table 1 shows a summary of statistical analyses.

Table 1

Statistical Analysis Plan

Hypothesis	IVs	DVs	Statistical Test	Effect Size
Sensitivity will not vary across prevalence conditions.	Prevalence rate (1%, 3%, 5%, 20%)	*d'* (parametric), area under the ROC curve (non-parametric)	Traditional ANOVA and Bayesian ANOVA	Partial eta-squared (η_p^2), Bayes factor (K), inverted
Criterion values will be higher in the low prevalence conditions than in the high prevalence condition.	Prevalence rate (1%, 3%, 5%, 20%)	*c*	One-way between-subjects ANOVA, planned comparisons	Partial eta-squared (η_p^2)
A main effect of prevalence on miss rate will be observed.	Prevalence rate (1%, 3%, 5%, 20%)	Miss rate	One-way between-subjects ANOVA	Partial eta-squared (η_p^2)

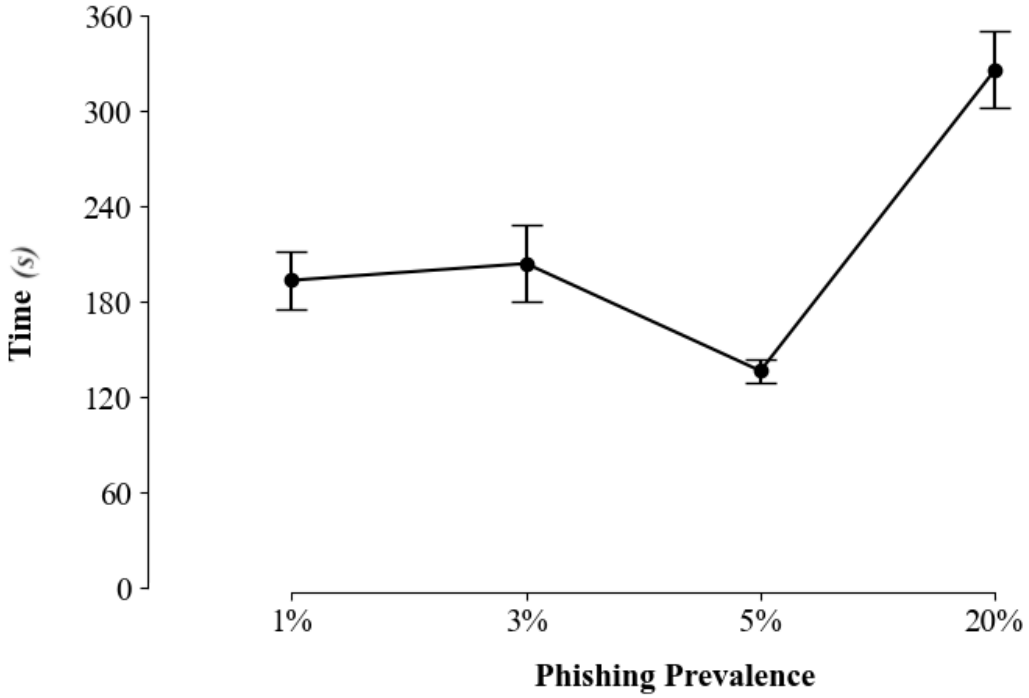
Results

Over 1 month, the survey collected 137 responses, of which 82 were included in analysis. Responses were excluded from analysis if they were left incomplete (less than 100% completion and inactive for more than 2 weeks) or if time spent on each page dropped unreasonably low, especially in comparison to earlier page completion times from the same participant. Thirty-two responses were excluded for incompleteness, while 23 responses were excluded for not completing the task appropriately. While it was possible for participants to start a survey and complete it in multiple sittings across two weeks, all participants—except two in the 1% and 3% conditions—did so in a single session. Twenty-one participants were in the 1%, 5%, and 20% conditions, while nineteen were in the 3% condition.

Time taken per page over all conditions averaged 191.69 seconds ($N = 82$, $SD = 348.16$), meaning participants spent on average 3.83 seconds per email. Participants in the 5% condition spent the fewest seconds per page ($M = 139.20$, $SD = 144.19$), averaging 2.78 seconds per email, followed by participants in the 1% condition, who spent 196.25 seconds on average ($SD = 370.58$), or 3.93 seconds per email. Participants in the 3% condition averaged 206.80 seconds per page ($SD = 472.5$), or 4.14 seconds per email. Participants in the 20% prevalence condition spent the most seconds per page ($M = 328.77$, $SD = 247.99$), averaging 6.58 seconds per email. Figure 3 graphs these means with standard error bars. This difference in time spent per page may be due to how the 20% condition showed less stimuli. Participants across the lower prevalence conditions began to spend less time per page as they progressed through the task, an informal observation that is in line with findings from an experiment on email load and detection accuracy by Sarno and Neider (2022).

Figure 3

Average Seconds Taken per Page of Fifty Emails per Phishing Prevalence Condition



Note. Error bars represent ± 1 standard error of the mean.

Hypothesis 1 - Sensitivity (*d'*)

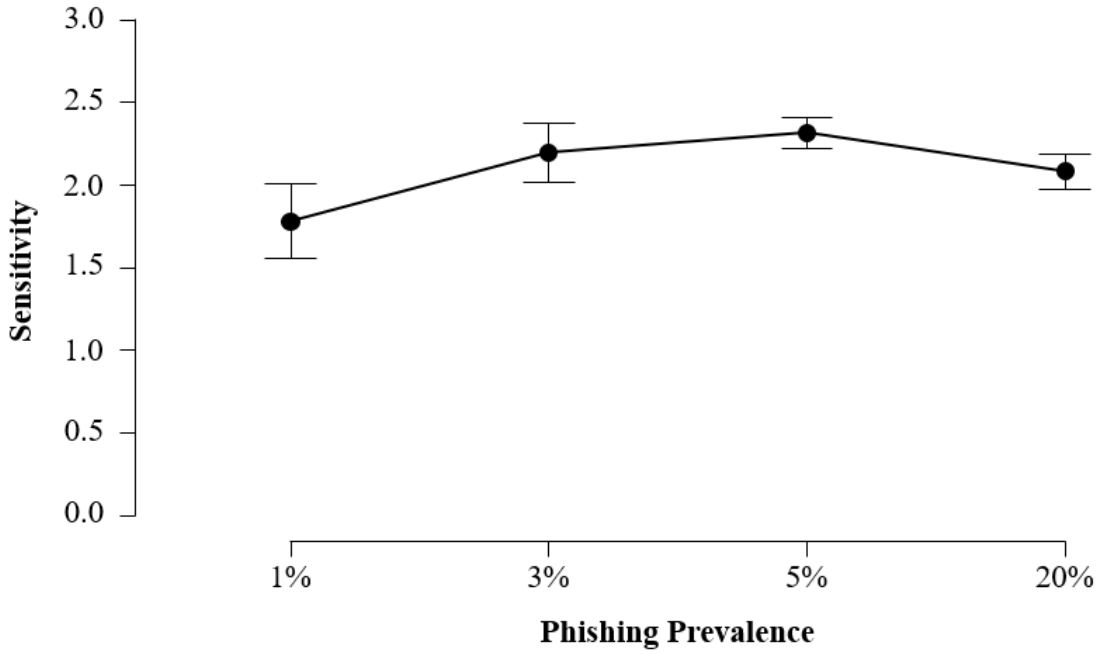
Mean sensitivity was lowest in the 1% condition (M = 1.78, SD = 1.03), followed by the 20% condition (M = 2.08, SD = .48), the 3% condition (M = 2.20, SD = .78), and the 5% condition (M = 2.32, SD = .44). Figure 4 provides a visual comparison of the means with standard error bars.

Prior to running an ANOVA to see if these values were significantly different, we tested assumptions of normality of distribution and equality of variance. Examination of a Q-Q plot showed an approximately normal distribution. However, Levene's test for equality of variance was significant, $F(3,78) = 6.73$, $p < .01$, indicating that the variances of sensitivity

were not equal across conditions. Thus, we conducted Welch's ANOVA, an alternative to the classic ANOVA used in cases of heteroskedasticity, which revealed no main effect of phishing prevalence on sensitivity, $F(3,41) = 1.97$, $p = .13$, $\eta_p^2 = .076$.

Figure 4

Average Sensitivity per Phishing Prevalence Condition



Note. Error bars represent ± 1 standard error of the mean.

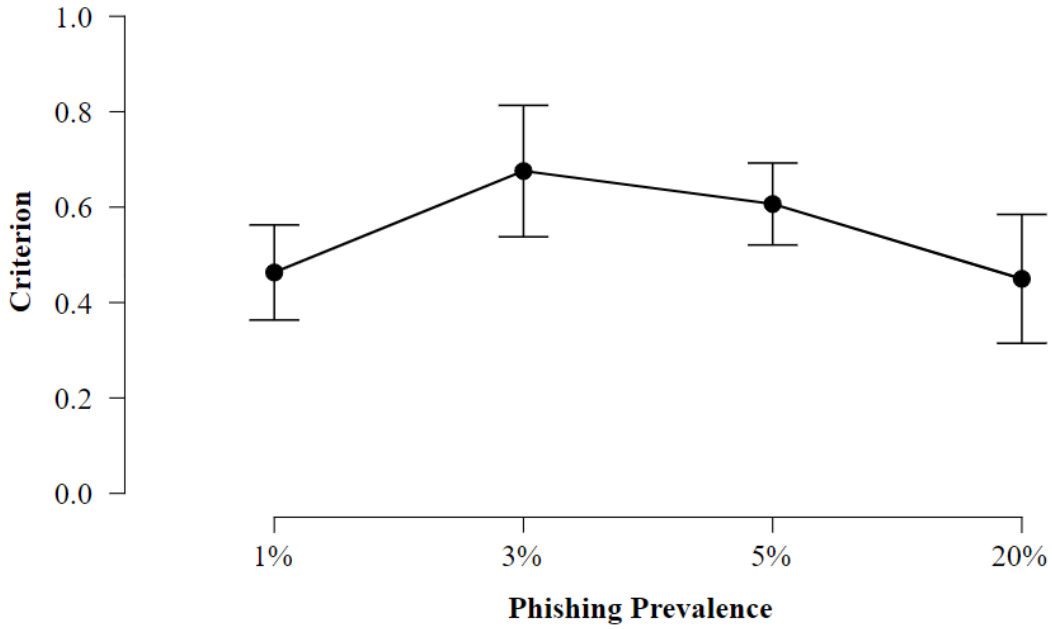
To examine if there was significant support for the null hypothesis, a one-way between-subjects Bayesian ANOVA was conducted to compare the effect of phishing email prevalence on phishing detection sensitivity in 1%, 3%, 5%, and 20% prevalence conditions. The Bayesian ANOVA revealed weak evidence supporting the null hypothesis, that phishing prevalence has no effect on detection sensitivity ($BF_{01} = 1.60$).

Hypothesis 2 - Criterion (*c*)

Mean criterion was highest for the 3% condition (M = 0.68, SD = 0.14), followed by the 5% condition (M = 0.61, SD = 0.09), 1% condition (M = 0.46, SD = 0.10), and 20% condition (M = 0.45. SD = 0.14). Figure 5 graphs average criterions and standard errors of each mean.

Figure 5

Average Criterion per Phishing Prevalence Condition



Note. Error bars represent ± 1 standard error of the mean.

Before conducting an ANOVA of phishing prevalence on criterion, we tested its assumptions for analysis. Examination of a Q-Q plot showed an approximately normal distribution. Levene's test, conducted to test the assumption of equality of variances, was not significant, suggesting that variances were equal across conditions, $F(3,78) = 1.18$, $p = .32$. A one-way between-subjects ANOVA revealed that differences in criterion between prevalence

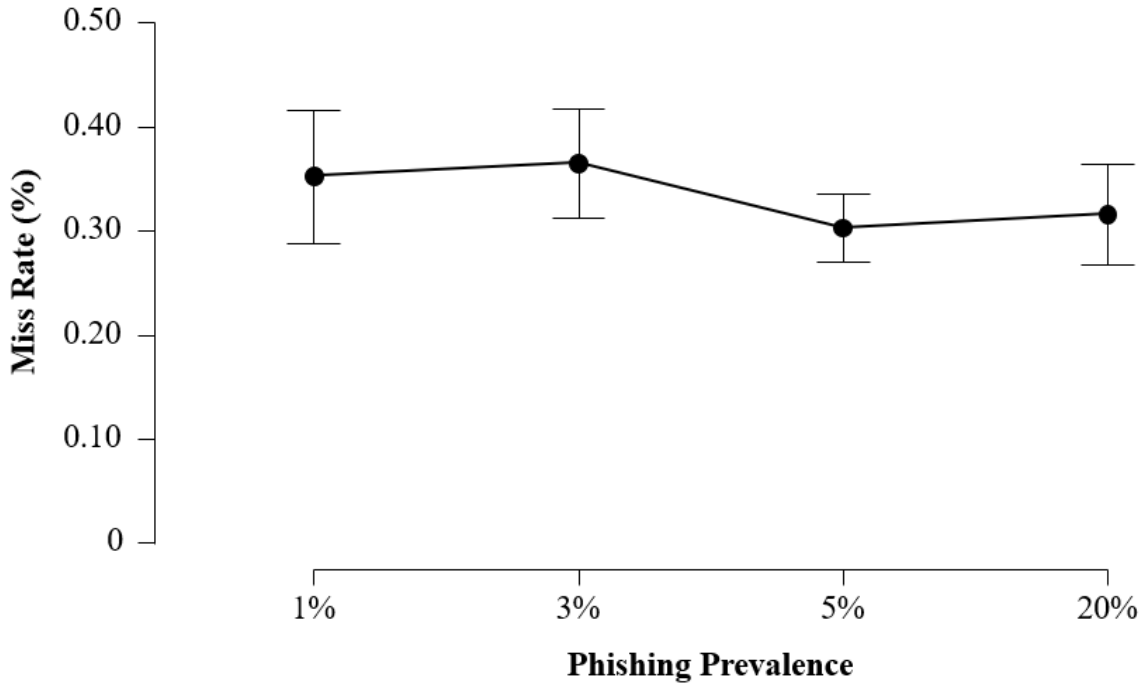
conditions were not statistically significant, $F(3,78) = 0.89$, $p = .45$, $\eta_p^2 = .03$. Because the ANOVA was not significant, planned comparisons were not done.

Hypothesis 3 - Miss Rates

Mean miss rate was highest in the 3% condition (M = 36.5%, SD = 22.9%), followed by the 1% condition (M = 35.2%, SD = 29.3%), the 5% condition (M = 30.3%, SD = 14.9%), and the 20% condition (M = 31.6%, SD = 22.1%). Figure 6 plots average miss rates and standard error per condition.

Figure 6

Average Miss Rates per Phishing Prevalence Condition



Note. Error bars represent ± 1 standard error of the mean.

Examination of a Q-Q plot showed an approximately normal distribution. However, Levene's test for equality of variance was significant, $F(3,78) = 4.19$, $p < .01$, meaning that

the variances of an arcsine transformation of miss rates were not equal across conditions. Thus, we conducted Welch's ANOVA, $F(3,41) = 0.18$, $p = .91$, $\eta_p^2 = .006$, and did not detect a significant main effect on miss rates.

Self-reported Prior Phishing Training

Forty-six out of 82, or 56% of participants, indicated that they have never had anti-phishing training prior to this study. Thirteen of these individuals were in the 1% and 20% conditions each, while the 5% condition had 15 and the 3% condition had just 5. Of the 36 who did have some prior anti-phishing training, 23 (or 64%) responded that they had training over a year ago.

Discussion

This study aimed to explore whether the low prevalence effect applies during phishing detection. To this end, four levels of prevalence were examined (1%, 3%, 5%, and 20%), with each of the low prevalence conditions including 1000 real-world emails and the high prevalence condition including 250 emails. To help manage such a large stimulus set and increase external validity, this study had participants look at images with 10 emails at a time, approximating an email inbox. The low prevalence effect is a phenomenon where miss rates increase when the prevalence of a target is low, but this increase is not accompanied by a change in sensitivity. We proposed three hypotheses: that sensitivity would not vary based on phishing prevalence, that criterion would increase as phishing prevalence decreased, and that miss rates would increase as phishing prevalence decreased.

Results of this study indicate weak or circumstantial support for the first hypothesis, that sensitivity would not vary based on phishing prevalence. We found neither strong evidence to suggest that meaningful differences in sensitivity existed, nor that sensitivity across conditions was the same. Further research will be needed to investigate any possible relationship between sensitivity and phishing prevalence.

These results partially replicated findings in similar emerging research. Sensitivity values for the 20% condition approximate those reported in the 25% and 50% conditions of Sarno and Neider (2022). Despite that, this study did not replicate their finding that sensitivity in the 5% condition was significantly lower than in either of the 25% and 50% conditions, while the latter conditions' sensitivities did not differ significantly from each other.

However, sensitivity values across all conditions ranged between 1.78 and 2.32, suggesting a good level of task difficulty for the set of stimuli used in this study. If the task was too difficult and participants were performing at near chance levels, a confound would have been introduced in that we would not have been sure that poor performance was due to the rarity of the target, or simply because the participants could not detect the target. On the other hand, if the task was too easy, and participants performed at or near 100% accuracy, we would not have had enough false alarms or misses to conduct signal detection analyses. The range of sensitivity values observed in this study indicates that the inbox simulation method, combined with creating random character strings for most phishing email sender addresses, may be a useful tool for future research using conduct signal detection analyses with a more varied, life-like stimulus set.

We observed no evidence in support of the second hypothesis, which predicted that participants in lower prevalence conditions would exhibit a positive criterion shift. This also replicates the finding reported by Sarno and Neider (2022), who did not find a main effect of prevalence on criterion. The criterion values we observed did replicate their observation that individuals were biased toward responding that emails were legitimate in ambiguous situations. However, the lack of a main effect on criterion contrasts with Sawyer and Hancock's (2018) research suggesting that the low prevalence is present during phishing detection, and other research suggesting that criterion shifts are responsible for the low prevalence effect (e.g., Wolfe & Van Wert, 2010). Despite this, we cannot definitively say that the low prevalence effect does not exist during phishing detection, as research into both

understanding the low prevalence effect and methods for better examining phishing detection are still emerging.

Finally, we also did not observe evidence in support of the third hypothesis, which predicted that miss rates would increase as prevalence decreased. These results did not replicate Sawyer and Hancock's results (2018), which found elevated miss rates in the 1% prevalence condition. However, our results did indicate that the low prevalence effect did not occur at 5% phishing prevalence, in line with existing research in the field (Sarno & Neider, 2022; Sawyer & Hancock, 2018).

A possible reason for not seeing the low prevalence effect in this study is that showing multiple stimuli at once allowed participants to revisit their responses, as long as they stayed on the same page. In addition, the fact that a legitimate email was matched with each phishing email could have served as a method of training individuals on the task as they completed it. Providing feedback and allowing individuals to correct errors have been shown to mitigate the low prevalence effect in certain visual search tasks (e.g., Fleck & Mitroff, 2007; Schwark et al., 2012). Thus, participants could have been given the information they needed to realize and fix errors from as far as 50 emails ago. Allowing individuals to correct errors and providing a possible form of training through the matched emails, combined with inserting a consistent characteristic to distinguish phishing emails from legitimate emails, may have mitigated the low prevalence effect.

While this study did employ a much larger stimulus set size to improve the validity of the research, several limitations and improvements to the study's design exist. The next section will discuss these and suggest directions for future research.

Limitations and Future Directions

Limitations of this study include the lightweight nature of the screener, which may not have adequately screened for English language fluency, a variable that a study has recently reported may negatively impact phishing detection performance (Hasegawa et al., 2022). Future studies could include more measures of reading speed or fluency to control for this. In addition, this study included a short training block in the interest of study length management, but future studies would benefit from a longer, more robust training block with more trials.

The lack of demographic information limits the reproducibility of results from this study as exact details of age and gender distribution are unclear. Future studies should collect demographics data such as gender or age even if they are not expected to impact the results of the study for the purposes of future analysis, attempts to reproduce the findings, or informing possible other directions of research.

Characteristics of this study's sample may have also contributed some limitations. The sample primarily included undergraduate students who aimed to get credit for a psychology class and who live in a geographical area (Silicon Valley) where knowledge of technology may be higher than other parts of the world. While a large proportion reported they had never received phishing training before this study, they may have been familiar with the concept at a level that differs from the general populace. Future research could be conducted on populations in other geographical areas and incorporate a more diverse range of ages.

The ten-stimuli per image method also introduced limitations. While it did better approximate the real-life situation of scrolling through an email inbox, the use of the

Qualtrics platform meant that it was not possible to examine exact time spent per email, nor collect information on how criterion or accuracy changed over time for each participant. The remote nature of the study additionally meant that we lacked insight into participant situation and behavior while performing the task. Future studies could include clarifying questions in the survey or be conducted in a way that enables gathering contextual information on what is happening when participants spend unusually low or high amounts of time on a page.

Low prevalence studies are difficult to conduct because they require a high volume of stimuli to ensure enough targets are shown to participants for analysis. For example, a study with 300 stimuli, which is already a large number of stimuli to source, would only provide three targets per participant in the 1% prevalence condition, meaning that analyses would be run on individual miss rates of 0%, 33%, 67%, or 100%. To date, studies exploring whether the low prevalence effect exists in phishing detection have used small sets of 300 stimuli or fewer (e.g., Sarno & Neider, 2020, Sawyer & Hancock, 2018; Singh et al., 2019) limiting the power of their results, while studies on tasks in other domains have utilized over 2000 (e.g., Evans et al., 2013; Wolfe et al., 2007). This study attempted to mitigate that by using a thousand emails sourced from the real-world. Future studies may be able to continue to refine this database of emails or the inbox simulation method used to display them.

Conclusion

This study aimed to examine the low prevalence effect in phishing detection using a simulated inbox scenario and, despite efforts to increase both internal and external validity compared to other research in this area, was unable to find evidence for the low prevalence effect in phishing detection. More research is needed to understand factors that impact

phishing detection. Phishing remains a huge and ever-adapting threat to individual- and business-level safety. While automated filters do exist and improve over time, humans remain the last line of defense against attackers who constantly employ new ways to steal identities, financial information, and more private information. Better understanding the root factors that impact the complex identification task will be instrumental for improving and creating increasingly effective phishing detection training strategies.

References

Abbasi, A., Zahedi, F. M., & Chen, Y. (2016). Phishing susceptibility: The good, the bad, and the ugly. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)* (pp. 169–174). Tucson, AZ, USA: IEEE. https://doi.org/10.1109/ISI.2016.7745462

Accenture. (2019). *Accenture's ninth annual cost of cybercrime study: Unlocking the value of improved cybersecurity protection.* Retrieved from https://mma.prnewswire.com/media/882924/Accenture_Cybercrime_Costs_Canadian_Companies_more_than_US_9M_La.pdf?p=original

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

Anti-Phishing Working Group. (2019). *2019 Q3 phishing activity trends report*. Anti-Phishing Working Group. https://docs.apwg.org/reports/apwg_trends_report_q3_2019.pdf

Beanland, V., Lenné, M. G., & Underwood, G. (2014). Safety in numbers: Target prevalence affects the detection of vehicles during simulated driving. *Attention, Perception, & Psychophysics*, *76*(3), 805–813. https://doi.org/10.3758/s13414-013-0603-1

Canfield, C. I., Fischhoff, B., & Davis, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 58*(8), 1158–1172. https://doi.org/10.1177/0018720816665025

Chen, Y., YeckehZaare, I., & Zhang, A. F. (2018). Real or bogus: Predicting susceptibility to phishing with economic experiments. *PLOS ONE, 13*(6), e0198213. https://doi.org/10.1371/journal.pone.0198213

Cofense Inc. (2017). *Enterprise phishing resiliency and defense report.* Retrieved from https://cofense.com/phishing-resiliency-report-2017/

Cofense Inc. (2019). *Cofense annual phishing report 2019.* Retrieved from https://cofense.com/wp-content/uploads/2019/10/phishing-report-2019.pdf

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00781

Downs, J. S., Holbrook, M. B., & Cranor, L. F. (2006). Decision strategies and susceptibility to phishing. *Proceedings of the Second Symposium on Usable Privacy and Security - SOUPS '06*. https://doi.org/10.1145/1143120.1143131

Egelman, S., Cranor, L. F., & Hong, J. (2008). You've been warned: An empirical study of the effectiveness of web browser phishing warnings. *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*, *1065*. https://doi.org/10.1145/1357054.1357219

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If you don't find it often, you often don't find it: Why some cancers are missed in breast cancer screening. *PLOS ONE*, *8*(5), e64366. https://doi.org/10.1371/journal.pone.0064366

Federal Bureau of Investigation. (2020). *2019 Internet Crime Report*. Retrieved August 25, 2020, from https://www.ic3.gov/Media/PDF/AnnualReport/2019_IC3Report.pdf

Federal Bureau of Investigation. (2022). *2021 Internet Crime Report*. Retrieved June 27, 2023, from https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf

Ferguson, A. J. (2005). Fostering e-mail security awareness: The West Point carronade. *Educause Quarterly*, *28*(1), 54–57.

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science*, *18*(11), 943–947. https://doi.org/10.1111/j.1467-9280.2007.02006.x

Hasegawa, A. A., Yamashita, N., Akiyama, M., & Mori, T. (2022). Experiences, behavioral tendencies, and concerns of non-native English speakers in identifying phishing emails. *Journal of Information Processing*, *30*, 841-858. https://doi.org/10.2197/ipsjjip.30.841

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00150

Hout, M. C., Walenchok, S. C., Goldinger, S. D., & Wolfe, J. M. (2015). Failures of perception in the low-prevalence effect: Evidence from active and passive visual search. *Journal of Experimental Psychology. Human Perception and Performance*, *41*(4), 977–994. https://doi.org/10.1037/xhp0000053

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, *50*(10), 94–100. https://doi.org/10.1145/1290958.1290968

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.2307/2291091

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*, *24*(2), 164–176. https://doi.org/10.1108/ICS-01-2016-0002

Kim, J.-H., & Jang, S. (2014). A scenario-based experiment and a field study: A comparative examination for service failure and recovery. *International Journal of Hospitality Management*, *41*, 125–132. https://doi.org/10.1016/j.ijhm.2014.05.004

Kleitman, S., Law, M. K. H., & Kay, J. (2018). It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. *PLOS ONE, 13*(10), e0205089. https://doi.org/10.1371/journal.pone.0205089

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007). Protecting people from phishing: The design and evaluation of an embedded training email system. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 905–914. https://doi.org/10.1145/1240624.1240760

Ly, A., Stefan, A., van Doorn, J., Dablander, F., van den Bergh, D., Sarafoglou, A., Kucharský, S., Derks, K., Gronau, Q. F., Raj, A., Boehm, U., van Kesteren, E. J., Hinne, M., Matzke, D., Marsman, M., & Wagenmakers, E. J. (2020). The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the p value hypothesis test. *Computational Brain & Behavior*, *3*(2), 153–161. https://doi.org/10.1007/s42113-019-00070-x

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

Martin, J., Dubé, C., & Coovert, M. D. (2018). Signal detection theory (SDT) is effective for modeling user behavior toward phishing and spear-phishing attacks. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *60*(8), 1179–1191. https://doi.org/10.1177/0018720818789818

Moreno-Fernández, M. M., Blanco, F., Garaizar, P., & Matute, H. (2017). Fishing for phishers. Improving Internet users' sensitivity to visual deception cues to prevent electronic fraud. *Computers in Human Behavior*, *69*, 421–436. https://doi.org/10.1016/j.chb.2016.12.044

Sarno, D. M., & Neider, M. B. (2022). So many phish, so little time: Exploring email task factors and phishing susceptibility. *Human Factors*, *64*(8), 1379-1403. https://doi.org/10.1002/acp.3594

Sawyer, B. D., & Hancock, P. A. (2018). Hacking the human: The prevalence paradox in cybersecurity. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *60*(5), 597–609. https://doi.org/10.1177/0018720818780472

Schwark, J., Sandry, J., MacDonald, J., & Dolgov, I. (2012). False feedback increases detection of low-prevalence targets in visual search. *Attention, Perception, & Psychophysics*, *74*(8), 1583–1589. https://doi.org/10.3758/s13414-012-0354-4

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

Siadati, H., Palka, S., Siegel, A., & McCoy, D. (2017). *Measuring the effectiveness of embedded phishing exercises*. 10th USENIX Workshop on Cyber Security Experimentation and Test (CSET 17). https://www.usenix.org/conference/cset17/workshop-program/presentation/siadatii

Singh, K., Aggarwal, P., Rajivan, P., & Gonzalez, C. (2019). Training to detect phishing emails: effects of the frequency of experienced phishing emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 453–457. https://doi.org/10.1177/1071181319631355

Van Wert, M. J., Horowitz, T. S., & Wolfe, J. M. (2009). Even in correctable search, some types of rare targets are frequently missed. *Attention, Perception, & Psychophysics*, *71*(3), 541–553. https://doi.org/10.3758/APP.71.3.541

Verizon. (2020). *2019 data breach investigations report.* Verizon Enterprise. Retrieved June 27, 2023, from https://www.verizon.com/business/resources/reports/dbir/2019/

Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, *51*(3), 576–586. https://doi.org/10.1016/j.dss.2011.03.002

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. https://doi.org/10.1016/j.chb.2017.03.002

Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of Vision*, *13*(3), 33–33. https://doi.org/10.1167/13.3.33

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches: Cognitive psychology. *Nature*, *435*(7041), 439–440. https://doi.org/10.1038/435439a

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General*, *136*(4), 623–638. https://doi.org/10.1037/0096-3445.136.4.623

Wolfe, J. M., & Van Wert, M. J. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology*, *20*(2), 121–124. https://doi.org/10.1016/j.cub.2009.11.066

Appendix

Qualtrics Training Block

In this study, you will be asked to pick out phishing emails from a set of emails as best as you can.

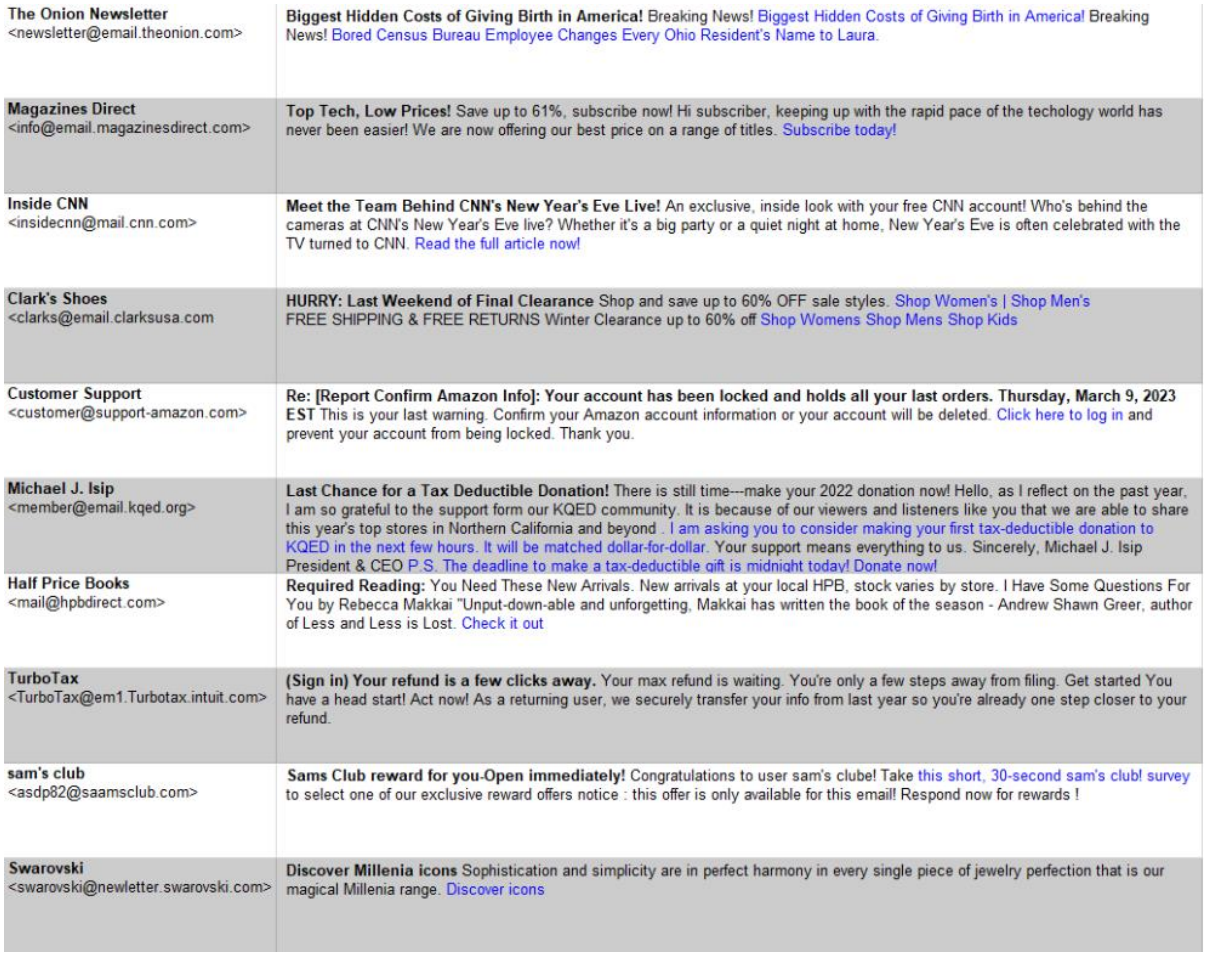
Phishing emails are emails from attackers who pretend to be someone else--a company/organization, a person you may know, or a fake identity--and try to get personal or login information from you. They may copy or imitate the colors and brand images of the entity.

Phishing emails look different from one to the next. They often include at least one of the following characteristics:

- a fake sender address
- a call for urgent action
- a generic greeting (e.g., Dear Valued Customer)
- grammatical and/or spelling errors
- a premise that can't be true

In the following practice task, you will be shown a list of email text previews similar to what you might see in an email inbox.

Please indicate which emails seem to be phishing emails in the following image by clicking on them to highlight them.



The Onion Newsletter <newsletter@email.theonion.com>	**Biggest Hidden Costs of Giving Birth in America!** Breaking News! Biggest Hidden Costs of Giving Birth in America! Breaking News! Bored Census Bureau Employee Changes Every Ohio Resident's Name to Laura.
Magazines Direct <info@email.magazinesdirect.com>	**Top Tech, Low Prices!** Save up to 61%, subscribe now! Hi subscriber, keeping up with the rapid pace of the techology world has never been easier! We are now offering our best price on a range of titles. Subscribe today!
Inside CNN <insidecnn@mail.cnn.com>	**Meet the Team Behind CNN's New Year's Eve Live!** An exclusive, inside look with your free CNN account! Who's behind the cameras at CNN's New Year's Eve live? Whether it's a big party or a quiet night at home, New Year's Eve is often celebrated with the TV turned to CNN. Read the full article now!
Clark's Shoes <clarks@email.clarksusa.com	**HURRY: Last Weekend of Final Clearance** Shop and save up to 60% OFF sale styles. Shop Women's | Shop Men's FREE SHIPPING & FREE RETURNS Winter Clearance up to 60% off Shop Womens Shop Mens Shop Kids
Customer Support <customer@support-amazon.com>	**Re: [Report Confirm Amazon Info]: Your account has been locked and holds all your last orders. Thursday, March 9, 2023 EST** This is your last warning. Confirm your Amazon account information or your account will be deleted. Click here to log in and prevent your account from being locked. Thank you.
Michael J. Isip <member@email.kqed.org>	**Last Chance for a Tax Deductible Donation!** There is still time---make your 2022 donation now! Hello, as I reflect on the past year, I am so grateful to the support form our KQED community. It is because of our viewers and listeners like you that we are able to share this year's top stores in Northern California and beyond . I am asking you to consider making your first tax-deductible donation to KQED in the next few hours. It will be matched dollar-for-dollar. Your support means everything to us. Sincerely, Michael J. Isip President & CEO P.S. The deadline to make a tax-deductible gift is midnight today! Donate now!
Half Price Books <mail@hpbdirect.com>	**Required Reading:** You Need These New Arrivals. New arrivals at your local HPB, stock varies by store. I Have Some Questions For You by Rebecca Makkai "Unput-down-able and unforgetting, Makkai has written the book of the season - Andrew Shawn Greer, author of Less and Less is Lost. Check it out
TurboTax <TurboTax@em1.Turbotax.intuit.com>	**(Sign in) Your refund is a few clicks away.** Your max refund is waiting. You're only a few steps away from filing. Get started You have a head start! Act now! As a returning user, we securely transfer your info from last year so you're already one step closer to your refund.
sam's club <asdp82@saamsclub.com>	**Sams Club reward for you-Open immediately!** Congratulations to user sam's clube! Take this short, 30-second sam's club! survey to select one of our exclusive reward offers notice : this offer is only available for this email! Respond now for rewards !
Swarovski <swarovski@newletter.swarovski.com>	**Discover Millenia icons** Sophistication and simplicity are in perfect harmony in every single piece of jewelry perfection that is our magical Millenia range. Discover icons

Boxed in red are the phish. The two cases here had multiple warning signs—grammar errors, fake sender addresses, a call for urgency, and vague premises--but some phish will have only a few.

The Onion Newsletter <newsletter@email.theonion.com>	**Biggest Hidden Costs of Giving Birth in America!** Breaking News! Biggest Hidden Costs of Giving Birth in America! Breaking News! Bored Census Bureau Employee Changes Every Ohio Resident's Name to Laura.
Magazines Direct <info@email.magazinesdirect.com>	**Top Tech, Low Prices!** Save up to 61%, subscribe now! Hi subscriber, keeping up with the rapid pace of the techology world has never been easier! We are now offering our best price on a range of titles. Subscribe today!
Inside CNN <insidecnn@mail.cnn.com>	**Meet the Team Behind CNN's New Year's Eve Live!** An exclusive, inside look with your free CNN account! Who's behind the cameras at CNN's New Year's Eve live? Whether it's a big party or a quiet night at home, New Year's Eve is often celebrated with the TV turned to CNN. Read the full article now!
Clark's Shoes <clarks@email.clarksusa.com	**HURRY: Last Weekend of Final Clearance** Shop and save up to 60% OFF sale styles. Shop Women's | Shop Men's FREE SHIPPING & FREE RETURNS Winter Clearance up to 60% off Shop Womens Shop Mens Shop Kids
Customer Support <customer@support-amazon.com>	**Re: [Report Confirm Amazon Info]: Your account has been locked and holds all your last orders. Thursday, March 9, 2023 EST** This is your last warning. Confirm your Amazon account information or your account will be deleted. Click here to log in and prevent your account from being locked. Thank you.
Michael J. Isip <member@email.kqed.org>	**Last Chance for a Tax Deductible Donation!** There is still time---make your 2022 donation now! Hello, as I reflect on the past year, I am so grateful to the support form our KQED community. It is because of our viewers and listeners like you that we are able to share this year's top stores in Northern California and beyond . I am asking you to consider making your first tax-deductible donation to KQED in the next few hours. It will be matched dollar-for-dollar. Your support means everything to us. Sincerely, Michael J. Isip President & CEO P.S. The deadline to make a tax-deductible gift is midnight today! Donate now!
Half Price Books <mail@hpbdirect.com>	**Required Reading:** You Need These New Arrivals. New arrivals at your local HPB, stock varies by store. I Have Some Questions For You by Rebecca Makkai "Unput-down-able and unforgetting, Makkai has written the book of the season - Andrew Shawn Greer, author of Less and Less is Lost. Check it out
TurboTax <TurboTax@em1.Turbotax.intuit.com>	**(Sign in) Your refund is a few clicks away.** Your max refund is waiting. You're only a few steps away from filing. Get started You have a head start! Act now! As a returning user, we securely transfer your info from last year so you're already one step closer to your refund.
sam's club <asdp82@saamsclub.com>	**Sams Club reward for you-Open immediately!** Congratulations to user sam's clube! Take this short, 30-second sam's club! survey to select one of our exclusive reward offers notice : this offer is only available for this email! Respond now for rewards !
Swarovski <swarovski@newletter.swarovski.com>	**Discover Millenia icons** Sophistication and simplicity are in perfect harmony in every single piece of jewelry perfection that is our magical Millenia range. Discover icons