Restricting the Link: Effects of Focused Attention and Time Delay on Phishing Warning Effectiveness

Justin Petelka
University of Washington
The Information School
Seattle, WA, USA
jpetelka@uw.edu

Benjamin Berens
Karlsruhe Institute of Technology
Applied Informatics and Formal Description Methods
Karlsruhe, Germany
benjamin.berens@kit.edu

Carlo Sugatan
University of Michigan
School of Information
Ann Arbor, MI, USA
carlosugatan@gmail.com

Melanie Volkamer
Karlsruhe Institute of Technology
Applied Informatics and Formal Description Methods
Karlsruhe, Germany
melanie.volkamer@kit.edu

Florian Schaub
University of Michigan
School of Information
Ann Arbor, MI, USA
fschaub@umich.edu

***Abstract*—Phishing warning researchers have proposed two forms of hyperlink restrictions for reducing phishing click-through rates: focused attention, which prevents users from proceeding to a suspicious URL until they click the uncovered link inside the warning; and time delay, which disables link clicking for a short period of time. Both measures aim to draw user attention to the warning and nudge them to carefully evaluate the respective link's URL. However, the effectiveness of these measures has so far not been comparatively evaluated. We conducted a mixed-methods online experiment (n=1,320) to understand differences in the effectiveness of focused attention and time delay both independently and together. Our study used an instrumented email inbox environment, in which participants were asked to assess emails and email hyperlinks. We found that, while both focused attention and time delay reduced click-through rates independently, the strength of these effects were significantly different from each other with focused attention being more effective than time delay. Combining both measures reduced CTR even further. We also found that participants who saw a warning with a time delay were more likely to hover over hyperlinks for longer than those who saw a focused attention warning. We discuss the implications of our findings for the design of anti-phishing warnings.**

1. Introduction

Phishing emails continue to be a prevalent and effective attack method in business, government, and personal contexts. While automated phishing detection approaches are improving, many cases still require human input to ensure that a flagged email is indeed a phishing attack and not a false positive. In these cases, users are shown a phishing warning. Prior work suggests that security notifications are most effective when they are shown just before a security hazard [1], [2]. For phishing warnings specifically, researchers have suggested phishing warnings that appear near suspicious links are more effective than warnings that appear as banner warnings above an email in the subject line or browser warnings that appear after a user has clicked a link [3]. Providing such link-centric warnings within an email creates space for users to analyze the email's content and the suspicious URL(s), and to engage with expert-based [4] and/or experienced-based [5] phishing and scam identification techniques.

In addition to studying the effects of a phishing warning's position, researchers have proposed two methods of restricting a user's interaction with a hyperlink in order to draw the user's attention to the warning: adding a time delay (i.e., blocking link interactions for some number of seconds [4]) and using focused attention (i.e., preventing email readers from proceeding to a suspicious URL until they click the true destination URL in the warning [3]). However, it is not yet clear how these approaches compare in their effectiveness. Differences in the effect of link restrictions on click-through-rate (CTR) could be leveraged to contextually apply link restrictions to phishing attacks with different risk levels.

To explore this, we conducted a mixed-methods 2x5 online experiment with 1,320 participants recruited on Prolific to investigate the effects of focused attention (on, off) and different time delays (0, 2, 3, 4, 5 seconds) on phishing warning effectiveness. We asked participants to evaluate the hyperlinks in 15 emails presented in an instrumented email inbox environment. For four emails (3 phish, 1 false positive) participants in treatment groups saw a link-centric phishing warning, the design of which was informed by a pre-study (see Section 3.1). We considered warning effectiveness holistically by measuring phishing click-through

rate, self-reported intrusiveness and helpfulness scores, and false positive click-through rate.

Our results show a number of significant differences between focused attention and time delay. On their own, focused attention was more effective than time delay at preventing people from clicking on links, but combining focused attention and time delay reduced CTR even more. However, this finding also applied to false positives, though participants who saw a time delay warning were slightly more likely to click on false positives and hovered for longer on suspected phishing links. Qualitative data suggests that time delays were sometimes welcomed, but were also seen as more restrictive than focused attention, particularly when participants believed the warning was a false positive. We conclude the paper by discussing the implications of our findings for phishing warning design, including matching different link restrictions to different levels of phishing risk and using link restrictions as methods for personalizing email security settings.

2. Related Work

Research efforts on human-centric phishing interventions primarily aims to assist users in effectively recognizing phishing attempts within their day-to-day digital interactions [2], [3], [4], [6], [7], [8]. Previous studies have explored the effectiveness of different phishing interventions, including examination of specific tools [4], [8], meta-analyses of existing work [9], longitudinal studies in large organizations [10], using telemetric data from web browsers [11] and, most relevant for our work, phishing warning design principles [3], [6], [7], [12].

2.1. Phishing warning design features

Security dialogues, warnings, and other awareness-raising measures have emerged as critical components in phishing interventions (in addition to education [13], training [14], [15], and design components [1], [16]). Where simulated phishing campaigns, a common organizational security practice, can have adverse effects on staff behavioral responses to phishing attacks [17], warnings can complement and improve organizational security strategies and resilience [10]. Researchers have examined phishing warnings in a variety of contexts, such as web browsers [11], [18], email clients [3], [19], and increasingly SMS texts [20], [21]. However email is the most studied medium for phishing attacks and anti-phishing warnings [10] since Business Email Compromise (BEC) continues to have substantial financial impacts on organizations around the world [22], [23].

In the context of email, different anti-phishing warning design features have been identified as important levers for shaping whether people click on phishing links, such as (inter)active vs passive warnings [3], [4], [6], [16], the location or placement of a warning relative to salient security indicators [3], [4], and the informational content of phishing warnings [19], [24]. Understanding how specific phishing warning features shape people's interactions with phishing emails and links helps us develop more effective anti-phishing strategies and instruments.

Researchers have proposed several ways to improve security warning design by modifying interactions with warnings, such as including "attractors" designed to draw attention to particular information or a specific region of the interface [1], [2], or delaying interactions to break people out of habitual actions [25], [26]. In prior efforts, subsets of the authors have independently integrated these findings into the context of phishing warnings through two different forms of hyperlink restriction, where the normal click interaction on a suspicious phishing link is altered or changed [2], [3], [4]: "focused attention" and "time delay."

In a previous study, Petelka et al. implemented a security interaction design principle called "forced attention" [3], which going forward we refer to as *focused attention*. Focused attention uses the idea of a "request attractor," or an interface modification designed to draw people's attention to salient information [2]. Focused attention disables clicking on a suspicious email link and displays the link's true destination URL (the salient information for a suspicious link) in a small popup warning. If a user wants to proceed they can do so only by clicking the URL in the warning message. This approach restricts people's interaction with a URL to within the confines of the warning message, encouraging them to scrutinize the link. Our initial research suggested that such focused interaction significantly improves the effectiveness of phishing warnings [3].

Similarly, *time delays* modify security dialog interaction by temporarily disabling or delaying interaction, encouraging more deliberate investigation of potential problems and breaking people out of habitual actions [19], [25], [26]. Prior work by Volkamer et al. evaluated time delays in the context of the TORPEDO anti-phishing warning add-on [19] and demonstrated the promise of using time delays in enhancing people's phishing detection. Despite the respective promises of these two link restriction techniques, it is not clear how they compare in terms of their usability and effectiveness against each other in the context of phishing emails and links. This paper compares the two forms of link restriction to better identify differences in their effect on people's ability to identify and avoid phishing links.

2.2. Anti-phishing security vectors

Besides a warning's design, it is imperative to consider phishing attack vectors to gauge the effectiveness of anti-phishing warnings. Email-based phishing campaigns vary in their appearance and method, including in message design (e.g., a simple text [27] versus an altered copy of a legitimate message [28]), the message's tone (e.g., imposing time pressure [29] or implying scarcity [30]), and obfuscation of URLs [15], [31], [32], [33], [34]. Consistent with prior studies [29], [35], our study focuses on evaluating the effects of different phishing warning features against different URL manipulation techniques. Researchers have proposed different ways to categorize how phishing attackers obfuscate malicious web domains in hyperlinks. For our study, we

include three different types of URL manipulations: gibberish domains [36], keyword-related domains [37], and brand-related domains [38]. Each of our participants saw one of each type of these phishing URLs.

Gibberish domains represent the simplest form of phishing URLs—the phishing URL has no contextual connection to the content of the email. Attackers often register these URLs using sequences of characters and numbers that appear random or at least pseudo-random (e.g., www.hrzzhfs.xyz/). With even minimal scrutiny of the URL, most users should be able to recognize these URLs as suspicious.

Keyword-related domains represent a more challenging type of manipulation to identify. URLs within this category incorporate words that align with the overall context of the email—for instance, www.client-mail-services.com/ in an email designed to mimic an official Gmail communication. However, these URLs typically omit any direct mention of the spoofed organization to evade domain registry scanning efforts by prominent organizations, thus misleading the recipient while superficially appearing legitimate.

Brand-related domains are the most sophisticated and difficult to detect URL manipulation. Here, the actual name of the organization is used within the domain portion of the URL, but it is subtly altered by adding trustworthy-seeming elements (e.g., mail.google-services.com/ instead of mail.google.com/). This type of phishing URL deceives participants into believing the URL is controlled by a legitimate organization, making it easier for phishers to lure unsuspecting users.

3. Phishing Warning Design Process

The design of our study's phishing warnings was informed by a pre-study to determine what text best conveys the potential risk of a suspicious link to participants. After discussing the pre-study, we provide an overview of our design of the focused attention and time delay warnings. All our warnings are link-centric, i.e., they are displayed over suspicious links (as opposed to a banner warning at the top of the email) and activate when a participant hovers over a suspicious link in accordance with findings in our prior study [3]. Our final warning designs share a similar base design (see Figure 1). We show the unmasked URL in the warning to make it easier for participants to assess the suspicious hyperlink. The warnings use consistent warning text, with minor changes based on whether the time delay restriction was active or inactive, and whether the warning utilized focused attention or not.

3.1. Warning Text Pre-Study

Prior work investigating phishing warning text, including our own works, used a variety of different warning texts [3], [4], [10]. To inform the text of our warning, we conducted a mixed-methods pre-study to understand how people interpret different wordings in the phishing warning. We recruited 485

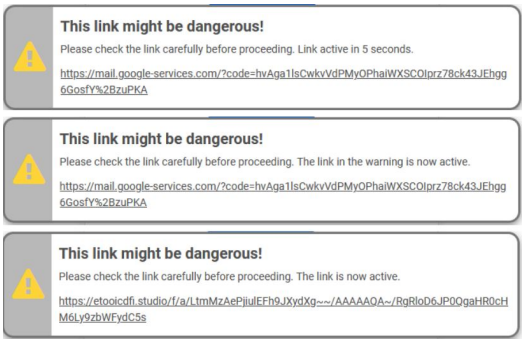

Figure 1. Our base warning design. The top image shows our warning when time delay is active. The second shows when time delay has elapsed and focused attention is active, i.e., only the link in the warning is clickable. The third shows when time delay has elapsed and focused attention is inactive, i.e., both the link in the email and in the warning are clickable.

participants from Prolific and showed them email screenshots, each of which contained a link-focused warning. We alternated each warning by changing the warning text's adjectives (e.g., unsafe, high-risk, dangerous, etc.) and subjects (e.g., website, URL, link, etc.) so that each participant saw different combinations of warning keywords (e.g., unsafe website). For each screenshot, we asked participants (1) what they thought the warning means (open-response), (2) how likely they would be to click on the link if they saw this message (5-point scale), and (3) how likely a number of outcomes would be if they clicked on the link (e.g., 'the website would load normally,' 'someone will steal my information,' etc.; 5-point scale per item).

We conducted ordinal logistic regression on likelihood to click responses and likely outcomes responses. Results from our ordinal logistic regression analysis found no significant main effect of keyword combinations on a participant's likelihood to click ($n.s.$), nor on their imagined outcomes ($n.s.$), i.e., all keyword combinations resulted in participants' not wanting to click the link.

In the absence of significant differences, we looked at the effect sizes of our logistic regression model, the Likert scale responses, and the open-ended responses to select a keyword combination. The goal was to identify a keyword combination that (1) decreased the likelihood to click on a suspicious link (i.e., a large negative effect size on likelihood to click a link), (2) that participants would interpret as phishing or stealing information as opposed to account suspension or a hacking attempt (i.e., accurate interpretation), and (3) that would convey the consequences of clicking a phishing link (i.e., accurate imagined outcomes). To code the open response questions, the second and third authors used thematic analysis [39] to develop and iterate a codebook over two rounds of analysis. In between rounds, the authors discussed their codes, identified discrepancies, developed consensus, and iterated the codebook. The final codebook was applied to all open responses and focused on themes such as the source of a warning, the accuracy of a participant's interpretation and what actions participants would take if they saw a similar warning in their inbox.

We found that the adjective "dangerous" led the most participants to correctly identify the URL as either trying to steal their information or as a phishing attack (i.e., our "correct interpretation" code) while also deterring people from clicking (i.e., a strong negative effect size). Similarly, we found that "link" may help to better focus participants on the risks of clicking a suspicious email link (as opposed to "web address" or "web page") and had a strong negative effect on likelihood to click. For these reasons we chose the text "This link might be dangerous!" as our warning header.

3.2. Link Restrictions

In our main study, we focused on two types of link restriction that are informed by prior work on security dialogues broadly and our own work on anti-phishing warnings specifically: time delay and focused attention.

3.2.1. Time delay. Time delay briefly disables clicking on suspicious links to break people out of habitual actions and give them time to assess potential security problems. For instance, the TORPEDO add-on prevents people from visiting a website for three seconds after they hover over a link [4]. While prior work has assessed the difference between having a time delay and not, there has not been an assessment on the effects of different durations of time delay. Therefore, we included five different levels of time delay: 0, 2, 3, 4, and 5 seconds. The zero-second delay served as a control condition (i.e., no time delay); 2-5 seconds as treatments. We internally piloted a one-second time delay, but decided to exclude it because it takes longer than one second to look at the warning and unmasked URL, resulting in no practical difference between 0 and 1 seconds of time delay.

3.2.2. Focused attention. Focused attention disables a suspicious link in an email entirely, instead only allowing them to click the unmasked URL presented in the warning. Preventing people from clicking on the link in an email body (whose URL may be masked, see Figure 2) is meant to explicitly focus a person's attention on the suspicious URL's true destination, which can help people identify and avoid suspicious hyperlinks [3]. We included two levels of focused attention in our treatment groups: warnings with focused attention and warnings without.

4. Methods

Through a 2x5 between-subjects online experiment (n=1,320), we examined the effects of time delay (0, 2, 3, 4, 5 seconds) and focused attention (yes, no). This included a baseline group with a phishing warning but without link restrictions (time delay: 0s; focused attention: no), four time delay-only conditions (time delay: 2, 3, 4, or 5s; focused attention: no), one focused attention-only condition (time delay: 0s; focused attention: yes), and four conditions combining both link restrictions (time delay: 2, 3, 4, or 5s; focused attention: yes). We further included an additional control group in which participants saw no phishing warning for eleven different groups in total.

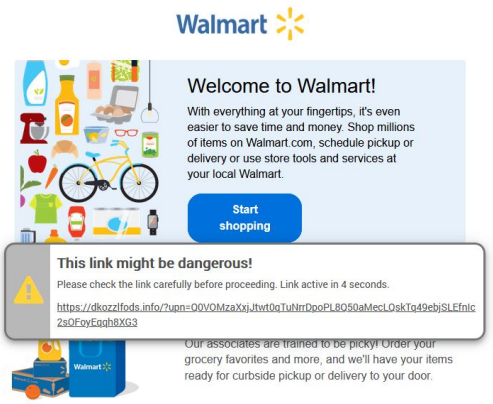

Figure 2. An example of how our warnings unmask the URL of a link. Here the full URL is shown in the warning, while the email hyperlink says "Start shopping."

Our study received approval from the University of Michigan's IRB and was pre-registered with the Open Science Foundation (OSF).[1]

4.1. Study Protocol

We recruited participants through Prolific and directed them to our survey on Qualtrics where they were randomly assigned to one of our eleven groups. We required participants to use a desktop device and excluded participants on tablet and mobile devices (checked in Prolific and Qualtrics) to ensure consistency of interactions with our inbox environment. Our survey and all other study materials are available in our OSF repository.[2] Our inbox environment code and analysis code is publicly available in our Github repository.[3]

We used deception to avoid priming participants for phishing risks. Our recruitment and consent form did not mention phishing, instead telling participants that their participation would help advance automated detection of inactive email hyperlinks. We discuss ethical considerations further in Section 4.5.

After agreeing to the consent form, we asked participants to view emails in an online email inbox, count the number of hyperlinks in each email, and evaluate whether links in these emails worked (which we defined as "leading to a working webpage"). We carefully designed this task to encourage participants to engage with links in emails without explicitly telling them to click on all hyperlinks. This framing placed the email evaluation task as the primary task, while identifying phishing links became a secondary task which replicates the task-switching required during actual phishing attacks [16].

1. OSF pre-registration is available at: https://osf.io/st6pz

2. Survey instrument and study materials are available in our OSF repository at: https://osf.io/chsv5/

3. Our inbox environment code is available at our Github repository: https://github.com/spilab-umich/phishing-experiment-infrastructure-2

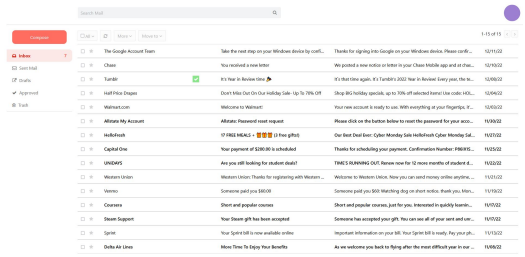

Figure 3. A picture of our main inbox screen.

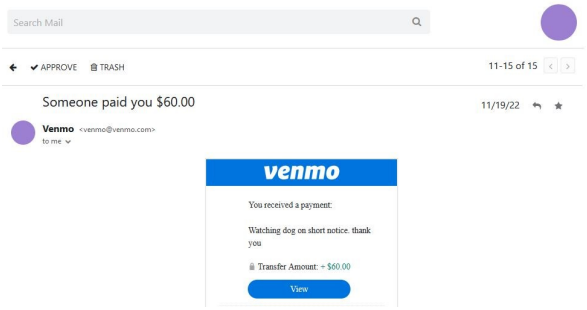

Figure 4. A picture of a single email view. We asked participants to assess whether the hyperlinks in each email were working. Participants indicated emails as working or not by clicking on the Approve or Trash buttons in the upper left corner of the email header bar.

Participants were given a link to our inbox and a unique username and password in Qualtrics. Asking participants to log in with individualized credentials was intended to make participants feel like they were logging into a peronal(ized) email account.

We based the style of our study's inbox environment (see Section 4.2) on Gmail (see Figure 3) to help our inbox feel intuitive and familiar to participants.

Once logged in, participants saw an inbox with 15 unread emails. Though we told participants that the set of emails was unique to them, each participant saw the same emails in random order (see Section 4.3 for email details).

In our instructions, we included a short list of things that might be "wrong" with a link as guidance for participants. We specifically included: links that do not load, have an error, are restricted, are suspicious, or have expired. We chose these heuristics to balance providing enough information so participants would understand and not be confused about the task without priming participants to the security focus of our study.

We asked participants to evaluate emails by labeling them using two buttons at the top of each email (see Figure 4). If participants found an issue with a hyperlink, we asked them to label the entire email as 'Trash;' if all links in the email appeared to work, we asked them to mark the email as 'Approved.'

Once participants labeled emails as Trash or Approved and input the number of working hyperlinks in each email into Qualtrics, we retrieved the total number of unread emails remaining in their inbox before allowing participants to proceed in our Qualtrics survey. If this number was not zero, participants were asked to return to their inbox and work through the remaining unread emails.

After completing the email evaluation, we asked participants follow-up questions in our survey. First, we asked participants in the treatment groups about their experience with the warning, including what they thought about the warning (open response), whether the warning was helpful (5-point Likert), whether the warning changed what they thought about the link (open response), whether the warning was intrusive (5-point Likert), and how the warning might be improved (open response). We then asked questions pertinent to the validity of our study: whether participants thought the warnings were a central part of the study, if participants noticed that some URLs were being re-directed, and whether participants clicked on links because they had assumed it was safe to do so in a study. Finally, we asked participants about their cybersecurity experience as measured by the SeBIS scale [40], their prior experience with phishing and scams, and their demographic backgrounds.

Once the survey was completed, we debriefed participants about the true purpose of the study and notified them that we recorded their email click and hover interactions. As suggested in prior work [41], [42], we also explained how phishing attacks spoof legitimate URLs and shared ways to identify or check phishing links to help participants avoid falling for future phishing attacks. After debriefing, we asked participants for consent a second time, and offered participants the option to opt out of the study without penalty as suggested by prior work [43]. After consenting a second time, we asked participants for optional feedback on their experience with the study (open response).

After completing the study, participants were provided a completion code to copy-and-paste back into Prolific. Participants were compensated $5 USD for work that was expected to take 20 minutes. The actual median completion time was 21:46 minutes, resulting in an average hourly rate of $13.76 USD.

4.2. Instrumented Email Inbox

Participants interacted with an online email client, shown in Figure 3, which we developed and instrumented for this study. Participants could log in and interact with emails, but some buttons and features (such as Compose) were visibly disabled. We modeled the design aesthetic after Gmail's web client to make the interface familiar and navigable. However, we did not include any name or logo to reduce brand effects (i.e., signaling this inbox was a Google-related product) which could artificially increase participant trust.

Our inbox recorded participant mouse interaction events (i.e., link clicks and hovers) on all hyperlinks inside of emails. Events were detected client side using jQuery and sent back to our web server using AJAX requests. Event records included participant's unique account id, the event's type (e.g., hover or click), timestamp, link id, and email id. We extensively tested our inbox environment internally and through pilot tests before running our experiment using

different combinations of browsers and operating systems. We also recorded each time a warning was displayed (i.e., when a participant actuates a warning by hovering) to assess potential changes in behavior after encountering multiple unique warnings. Using these "warning displayed" event records, we labeled participant data into time segments separated by how many warnings a participant had seen up to that point.

4.3. Email Selection and Phishing Emails

All participants saw the same fifteen emails in random order. These emails were altered versions of real emails the authors had received from a diverse range of companies and organizations. We removed hyperlinks that may have been difficult to see or links nested within each other to facilitate the primary task of identifying and analyzing email hyperlinks. We sanitized links and emails to remove any personally identifying information about the authors. This included removing references to personal email addresses, adjusting hyperlinks that changed an account's settings (e.g., Unsubscribe links), and anonymizing email text (e.g., for an email from the payment provider Venmo we changed the title from "<Author's Name> sent you a payment." to "Someone sent you a payment.") Finally, we removed or obfuscated email tracking methods to protect participant privacy, including modifying links with tracking query strings and deleting tracking pixels and images. The final modified versions of the 15 emails contained 3–11 links with an average of 5.5 links per email. Each email had at least one prominent call-to-action link (e.g., a "Pay Now" button, or "check the status of your order" link).

4.3.1. Phishing emails. We chose three of the fifteen emails to be our phishing emails. We chose these three emails because they came from organizations that are widely recognized in the US (i.e., Western Union, Google, Walmart). These three emails also all contained clear and distinct call-to-action links, which are common in phishing campaigns [44], [45], [46]. In these emails, we replaced the URL of the call-to-action hyperlink with a phishing URL that corresponds to one of three manipulation strategies (see Section 4.4).

In order to have participants safely interact with phishing links in our study, and to ensure that we measured the effect of our phishing warnings rather than aspects of a loaded website (e.g., phishing URLs prominently shown in browser address bar after clicking one of our phishing links), we implemented a solution that (a) displayed a real phishing link in the respective email, (b) revealed the real phishing URL in the status bar when hovered over, but (c) redirected participants to a safe website if they clicked the phishing link (e.g., clicking on the phishing link westernunion-pay.com would forward to the actual website westernunion.com). To accomplish these three goals, we used a clickjacking method. We disabled a link's "clickability" by hard-coding the link's onclick function to return false. We then added a transparent HTML element over the disabled phishing hyperlink. Thus, when a participant hovered over a phishing link, the phishing URL is displayed in the browser's status bar (and for the treatment groups a phishing warning is shown). When clicking the link, participants instead clicked the transparent HTML element with its own onclick event, which forwarded participants to a safe website. We discuss the ethical considerations of this approach in Section 4.5 and the limitations of this approach in Section 4.8.

4.3.2. False positives. We included one false positive email, i.e., a benign email that incorrectly displayed a phishing warning, to study whether our phishing warnings served their function of helping participants assess a suspicious link or whether participants simply adhered to the warning without further scrutiny. To do this, we randomly selected one of the twelve benign emails to serve as a false positive email for each participant. For the selected false positive email, we used our same approach with phishing emails, i.e., selecting a call-to-action link for the warning to be displayed, but we did not replace the URL with a phishing link. While the false positive email was selected randomly, its position was fixed as the second-to-last email in the participant's inbox to ensure that participants in treatment groups would have seen all three true positive phishing warnings before the false positive.

4.4. Phishing URL Types

We assessed the effectiveness of focused attention and time delay link restrictions against different types of phishing URLs. Prior work has shown that some forms of domain manipulation are easier for users to spot than others [47], [48]. For instance, people have difficulty differentiating between a company's name in a URL's domain or subdomain [47], [49]. Typosquatting attacks are also difficult for users to spot [50].

We also looked at reported phishing domains on PhishTank [51] to inform the phishing URLs we used for each type, and found that many PhishTank URLs were comprised of a random string of letters or otherwise made no attempt to spoof an organization's legitimate domain. From these, we selected the following three types of URL manipulation:

Gibberish domains. URLs that do not look like an authentic domain, e.g. hejkdsakda.xyz. We hypothesized that, given the high frequency of gibberish domains reported as phish on PhishTank, this URL manipulation would be easier to spot than other URL types (i.e., lower CTR), and best addressed by link-focused warnings with low time delays.

Keyword-related domains. URLs that contain words related to a service but not the brand's name, such as mail-client-services.com when spoofing gmail.com. We hypothesized this would be easier to detect than brand-related typo domains but harder than gibberish domains.

Brand-related domains. URLs that spoof a legitimate domain by including the brand's name in the domain name or

TABLE 1. PHISHING DOMAINS USED IN OUR STUDY.

Email	URL Type	URL
Google	Gibberish	https://www.hrzzhfs.xyz/?dU=v0G4RBKTXg2Gtk9jdyT5C0QhB-NuuHcbnI3N3H6KuOOlwYtyYUs_03KA==&F=v0fUYv
	Keyword	https://www.client-mail-services.com/_t/c/A1020005-1735F31E6028AC6D-68C618EC?l=AABkT3mkCxlWQlg7
	Brand	https://mail.google-services.com/?code=hvAga1lsCwkvVdPMyOPhaiWXSCOlprz78ck43JEhgg6GosfY%2BzuPKA
Western Union	Gibberish	https://dkozzlfods.info/?upn=Q0VOMzaXxjJtwt0qTuNrrDpoPL8Q50aMecLQskTq49ebjSLEfnlc2sOFoyEqqh8XG3
	Keyword	https://www.financial-pay.info/global-service/?upn=9O-2F0uOvVudG71uY6JZBiNBA2kJ1h0T8XTI4yLNm5Md
	Brand	https://www.westernunion-pay.com/global-service/track-transfer/?mid=IDS2303139625717 4xZOq8beIND
Walmart	Gibberish	https://etooicdfi.studio/f/a/LtmMzAePjiuIEFh9JXydXg~~/AAAAAQA~/RgRloD6JP0QgaHR0cHM6Ly9zbWFydC5s
	Keyword	https://www.online-shopping-payment.com/?ctPayload=H4sIAAAAAAAAA12QwW7CMBBE22F8VnKtmxHSc5FrXqsQ
	Brand	https://www.walmart-payment.com/?upn=31lcBBFrKkrK4MwiV2J2egimukuh7R5G2XSsnoDDvoYMcZXguaG-2BaZjU

subdomain, such as google-mail-services.com. We hypothesized that, in line with prior work [52], this would be the most difficult domain manipulation to spot (highest CTR).

Each participant saw all three types of domain manipulation in randomized order with each of the three phishing emails using a different type of URL manipulation. To accomplish this, we created three different phishing URLs for each email (see Table 1). All phishing URLs had the same total length (95 characters) to ensure that they would appear consistently in the phishing warnings in our experiment.

4.5. Study Ethics

While our study was approved by the University of Michigan's institutional review board (IRB), there are limitations in relying solely on IRBs for ethical guidance. We articulate our ethical considerations for our study here.

We carefully designed our study to ensure that interactions with phishing emails was realistic yet did not actually put participants at risk. Phishing links in our emails redirected to legitimate websites when clicked. We also verified that all our phishing links were not registered / leading to actual websites both before and during the study. We further disabled alternate methods of opening hyperlinks, such as keyboard shortcuts (e.g., CTRL + left click) and context menus (e.g., Open Link in New Window) to ensure participants would not travel to potentially malicious websites.

Our study involved deception in that we did not initially disclose the true purpose of the study (study effectiveness of phishing warnings) and that we recorded click and hover actions. While our use of deception was IRB approved, we also followed best practices for deception in experimental research, including disclosing the possibility of deception in our consent form (also referred to as "authorized deception" [42] or "forewarning" [53]). During debriefing, we not only disclosed our study's true intent but explained how phishers might disguise URLs to seem legitimate to help participants learn how to spot real phishing attacks [41], [42]. Last, we had participants re-affirm consent after the debriefing and gave them the option to opt out of the study without penalty if they felt uncomfortable with our use of deception [43]—we removed five such participants from our data.

Finally, we borrowed from clickjacking strategies to simulate risk while participants evaluated emails in our inbox. Our study design prevents the need for registering and controlling a real domain name that could be used for phishing, which allowed us to realistically spoof particular company domains (i.e., Walmart, Western Union, and Google) in our experiment without activating organizational security teams that monitor for domain squatting and abuse. While clickjacking is typically used for nefarious purposes, we used it to simulate risk for participants without actually putting them at risk [54].

4.6. Recruitment and Participants

We recruited a total of 1,380 participants via Prolific. Five participants withdrew their consent after debriefing and were excluded. To create a consistent base for comparison of click through rate between warnings, we excluded 55 participants who did not hover over all three phishing links. The remaining 1,320 participants interacted with all three phishing links, but 19 of them did not hover over their false positive link. We include these 19 responses in our analysis of phishing links as the three phishing emails always appeared before the false positive, but excluded them when analyzing false positive links. This left us with 1,320 participants for phishing link analysis (with 117–124 participants per group) and 1,301 participants for false positive analysis.

Our participant sample was fairly representative of the US population, though skewed towards being young and well-educated. Participants were 18–98 years old (Median: 32 years old). Our sample was relatively gender balanced: 636 (48.2%) people identified as men, 628 as women (47.6%), 40 as non-binary or trans (3.0%), 16 chose not to say (1.2%). Most identified as White or Caucasian (857, 65%), followed by Black or African American (136, 10.3%), Asian (incl. South and Southeast Asian; 100, 7.6%), Latin or South/Central American (84, 6.4%), Middle Eastern or North African (5, 0.4%), or American Indian or Alaskan Native (2, 0.2%); 116 (8.8%) reported mixed heritage; 20 (1.5%) preferred not to report their racial or ethnic identity. Participants were relatively educated, with 193 (14.6%) participants holding a graduate degree and 663 (50.2%) a Bachelor's degree; 451 (34.2%) had not obtained a Bachelor's degree, and 13 (0.98%) preferred not to say. For income, 397 (30.0%) participants reporting household income above $80,000 (above the 2022 US Census Bureau's median household income [55]), 469 (35.5%) participants reported income of $40–80,000, and 373 (28.2%) reporting less than $40,000 per year (near the US Census Bureau's poverty line for families of four [56]). 81 (6.1%) participants preferred not to report their household income.

4.7. Hypotheses and Data Analysis

Our data analysis focused primarily on comparing the effectiveness of phishing warnings that use focused attention and time delay as link restrictions, for which we formulated five pre-registered hypotheses.[1] We further conducted exploratory analyses (also described in the pre-registration), including participants' interactions with false positives and different phishing URL manipulations, mixed-effect regressions on predictive factors for clicking on different types of link, and qualitative analysis of open-text responses.

Effects of focused attention and time delay. We formulated five hypotheses around the effect of warnings and link restrictions for this study. First, we hypothesized that people are less likely to click on a suspicious link when presented with a warning [3], [6], [18]. Evaluating whether the presence of a warning reduces click through rates compared to those who saw no warning (control) is both an affirmation of prior work and serves as a baseline for our study (i.e., there may be an issue with our study design if we did not observe a difference in phishing CTR between treatment and control groups):

- H1: Phishing CTR will be significantly higher for the control group (no warning) than the treatment groups.

For time delay, prior work suggests that time delays help people break out of habitual actions and scrutinize security cues like suspicious phishing URLs [4], [25], [26]. On the other hand, time delays that are too long may annoy people. Thus, our study evaluates two hypotheses for time delay:

- H2a: Warnings using higher time delays will have significantly lower phishing click through rates.
- H2b: Warnings using longer time delays will have significantly higher values of self-reported intrusiveness.

For focused attention, prior work suggests that drawing participant attention to a hyperlink's URL destination helps people identify and avoid phishing links [2], [3]. We expect that the focused attention strategy does not create an obstacle that's perceived as intrusive by participants. As such, we expect that:

- H3a: Warnings that use focused attention will have significantly lower CTRs than warnings without focused attention.
- H3b: Warnings that use focused attention will not have different self-reported Intrusiveness levels than those that use focused attention.

We tested our hypotheses using omnibus tests (i.e., Mann-Whitney U-tests and Kruskal-Wallis) with posthoc comparison tests (i.e., Dunn tests).

False positives. We conducted exploratory analysis (i.e., no fixed hypothesis) of interactions with false positive emails in a similar way (Kruskal-Wallis followed by Dunn comparison tests).

Phishing URL types. To explore differences between focused attention and time delay on phishing click-through rates for different phishing URL types, we grouped our data by our three types of phishing URLs and conducted Mann-Whitney U tests to help us understand how phishing warning effectiveness for each of the three different types of phishing URL differs with warnings that use focused attention and time delay.

Exploratory regression analyses. We complemented our comparison tests with exploratory mixed-effect regression models to further contextualize our findings. We conducted ordinal logistic regressions for phishing clicks using focused attention and time delay as both independent fixed effects and as an interaction effect. We also examined the predictive effect of different amounts of time delay on phishing clicks. For phishing links, we also included the type of phishing URL (i.e., gibberish, brand, keyword) as a fixed effect in our model. We conducted similar analyses for false positive and benign links. In addition, we conducted a linear regression on hover time for benign links. This helped us develop a broad picture of whether and how focused attention and time delay shaped how participants interacted with links overall.

Qualitative analysis. We conducted qualitative analysis on participants' open responses regarding (1) how they perceived the warning, (2) whether the warning changed what they thought about the suspicious hyperlink, (3) how they would improve the warning, and (4) study feedback. We used a thematic and iterative coding strategy [39]. The first author looked through half the open-ended responses to develop an initial codebook. The first and last authors discussed the emerging themes and iterated the codebook, then the first author coded the remaining responses. After a second round of iteration, the first author reviewed the coded data to ensure validity of identified themes. The identified themes centered on how participants assessed suspicious links (Assessment), usability aspects of the warnings (Usability), effects of our study on participant behavior (Study Validity), and a code for when participants mentioned learning something from our study (Education).

4.8. Limitations

Our participant sample was fairly diverse but skewed towards younger affluent people and high levels of educational attainment (64.6% had a Bachelor's degree or higher). In free response questions, several participants mentioned having to take phishing training as part of their office job. As a result, some of our participants may have been more aware of phishing attacks and counter measures than the general US population.

At the same time, 19 participants mentioned in open response questions that the presence of warnings in a hyperlink clicking task gave them pause to wonder about the study's intentions or led them to consider withdrawing from the study. Some of these participants were concerned that we might be harvesting auto-populated credentials by asking them to click on links and open websites, even if the links were benign. Still others were worried that we were thoughtlessly exposing participants to risk by not first checking emails to ensure they were safe. This suggests our data may be slightly skewed towards people with high confidence in the vetting done by our institution and Prolific.

Next, we discuss the ecological validity of our study design and the impact of our design choices on participant behavior. We deliberately chose to redirect participants who

clicked on our "phishing" links to the relevant legitimate website. As a result, 131 participants mentioned reduced trust in the warning's accuracy and/or clicking through subsequent warnings since some or all of the warnings landed on legitimate websites. However, directing participants to either a second warning or a simulated phishing website would have likely had the opposite effect, causing participants to overtrust our phishing warnings, in addition to introducing confounding factors into the experiment.

It is also unclear how realistically participants responded to our study and how well we were able to obfuscate our study's true intention. A majority of participants responded that they both believed the warnings were a central part of the study (71%) and that they clicked on links because they felt safe to do so in a research study (59%). In our free response questions, 36 participants mentioned that they clicked on links they may not have in real life. However, we also have indications that participants engaged in subjective and realistic sensemaking in whether or not to click on a suspicious link. Not only did some participants click on links and others did not, but participants made subjective assessments in their free response questions, like P2888 who said: "since the [URL] had both Western Union in the address and the address started with an 'https,' I was confident that it was going to lead where I thought it would." Notably, the domain for their Western Union phishing link was "westernunion-pay.com" not the legitimate "westernunion.com."

Our qualitative data also suggests that participants engaged with subjective sensemaking when deciding how to complete our task and that they identified phishing links to the best of their abilities. There were also no significant differences in participant CTR on benign call-to-action links (i.e., the number of people likely to click on benign links was balanced between groups, see Table 2), nor any differences in benign link hover times between groups.

Taking all of this into account, we are not claiming that the specific link interaction rates we observed would be replicated exactly in real-world email settings. Our observed rates are likely overestimates (i.e., higher link click rates) of how participants might interact with emails in their own inboxes. However, we are confident that our study has high internal validity, and that the different link restrictions (i.e., time delay and focused attention) would have effects on anti-phishing effectiveness and CTR in real-world settings that would be in line with the differences observed in our study.

We limited our participants to desktop or laptop devices and excluded mobile users, even though people often interact with emails on smartphones and tablets. We did not include touchscreens to make our experimental design tractable. The method of interacting with hyperlinks on touchscreens are touches instead of clicks and hovers, which would have created a further additional factor in our experiment, as would the use of a mobile-responsive layout. Additional work is needed to examine anti-phishing warnings in touchscreen environments.

While email remains a prevalent channel for phishing attacks [22], phishers are increasingly using SMS text messages, social media platforms, and a combination of different channels to scam and trick people [23]. Our results may not extend into these other channels, and more research is needed for supporting people in detecting phishing attempts in non-email contexts.

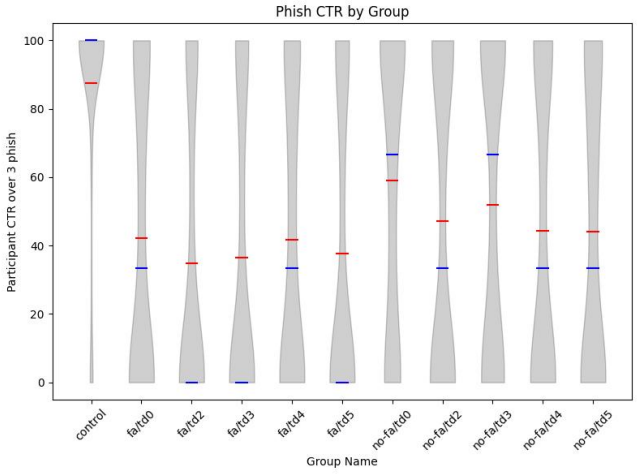

Figure 5. Phishing link click-through rate (CTR) violin plots for each of our 11 conditions, with median (*blue line*) and mean (*red line*) CTR values. The varying width represents the number of participants for each CTR value. "fa" stands for "focused attention", and "td" stands for "time delay" (e.g., the fa/td-2 group's warning used focused attention and a 2 second time delay).

5. Results

In this section, we first provide a descriptive analysis of participant interactions with links in our study, then discuss the effects of focused attention and time delay on phishing warning effectiveness, the effectiveness of warnings against different phishing URL types, and report our qualitative insights. We found that while both focused attention and time delay are effective at preventing people from clicking phishing links, focused attention was more effective than time delay. We also found that time delay was more noticeable than focused attention in open-ended response questions.

5.1. Participant Link Interactions

First we examine participants' overall interaction behaviors with phishing, false positive, and benign links. For phishing links, participants who did not see a warning (control) had much higher phishing click-through rates (CTR) than participants in the treatment groups (see Figure 5). We also see that groups with focused attention (FA) had slightly lower phishing click through rates than those without.

We see a similar pattern for false positive CTR (see Figure 6). The control group was much more likely to click on their false positive than those who saw a warning. We also see that focused attention groups have lower click-through rates than non-focused attention groups, which suggests that

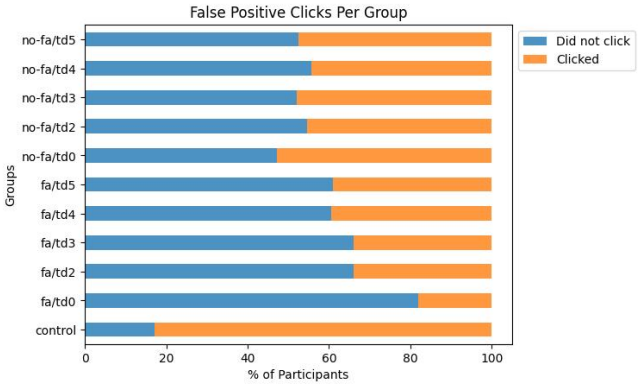

Figure 6. False positive CTR proportional bar chart for each of our 11 groups, showing the percent of each group that did or did not click on their false positive link.

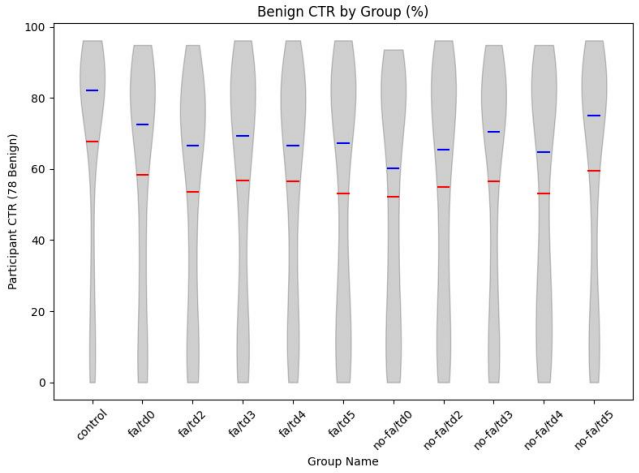

Figure 7. Benign CTR violin plots for each of our 11 groups, with median (*blue line*) and mean (*red line*) CTR.

the effect of focused attention is consistent for both phish and false positives.

For benign links (see Figure 7), we see that the control group has higher median CTR than those that saw a warning, but the effect is less pronounced than phishing and false positive CTR. This suggests that participants engaged with benign links consistently across groups, though it does appear that participants who saw a phishing warning were more hesitant to click on links overall than the control (no warning) group. We also see that benign CTRs are relatively consistent across all treatment (warning) groups suggesting that, while focused attention led participants to click less often on phishing links, it did not lead participants to avoid clicking on benign links; just links which displayed a phishing warning.

Benign CTRs appear relatively low considering the task was to assess whether hyperlinks are working or not. However, click-through and hover rates for call-to-action links, i.e., the most prominent link in each email, are substantially higher as shown in Table 2. While overall benign CTRs are between 53–67%, these rates greatly increase for links that explicitly call participants' attention (74–85%). We also found no significant differences in CTR for benign call-to-action links among groups (Kruskal-Wallis: n.s.).

TABLE 2. MEAN CLICK (CTR) AND HOVER (HOV) RATES FOR ALL BENIGN AND CALL-TO-ACTION (CTA) LINKS ONLY

Group	CTR Benign	CTA	Diff	HOV Benign	CTA	Diff
Control	67.62%	85.31%	+17.69%	89.61%	98.83%	+9.22%
TD0	52.14%	74.38%	+22.24%	86.45%	98.50%	+12.05%
TD2	55.03%	76.62%	+21.59%	88.45%	98.55%	+10.1%
TD3	56.49%	76.32%	+19.83%	89.61%	98.93%	+9.32%
TD4	53.11%	78.24%	+25.13%	85.66%	98.14%	+12.48%
TD5	59.55%	79.20%	+19.65%	89.24%	98.77%	+9.52%
FA	58.41%	75.86%	+17.45%	89.46%	99.25%	+9.79%
FA-TD2	53.53%	77.46%	+23.93%	86.85%	99.02%	+12.17%
FA-TD3	56.77%	78.67%	+21.89%	87.69%	98.90%	+11.21%
FA-TD4	56.52%	80.60%	+24.08%	86.43%	98.24%	+11.82%
FA-TD5	53.01%	75.50%	+22.49%	89.15%	98.07%	+ 8.93%

Hover rates are high for all benign links (85–89%) and higher for benign call-to-action links (98–99%). This suggests that participants actively engaged with links in our emails, but did not necessarily feel the need to click on all links to assess whether the link "worked," again suggesting participants engaged in subjective sensemaking as they might in the real world, rather than simply clicking on all links.

5.2. Focused Attention and Time Delay

5.2.1. Hypotheses testing. Next we report the results of our statistical tests for our hypotheses.

Phishing CTR significantly higher without a warning (H1). A Kruskal-Wallis test on phishing click-through rate identified significant differences among groups (χ^2=0.07, p<0.01) and a follow-up Dunn comparison test found significant differences between the control group and each of the treatment groups (see Table 3). The average phishing CTR for the control group (no warning) was 87.5%, compared to 43.9% for treatment (warning) groups. This suggests that participants who did not see a warning (control group) were significantly more likely to click on phishing links. Consistent with similar findings in prior work, this confirms H1: *Phishing CTR is significantly higher for the control group (no warning) than the treatment groups.*

Phishing CTR lower with time delay (H2a). We then examined the 1,203 participants who saw a warning to examine differences between focused attention and time delay as warning features. Comparing phishing CTR between participants whose warnings did and did not use a time delay, a Mann-Whitney U-test found significant differences (r=−0.07436, p=.001) between groups that had a time delay (average phish CTR: 42.2%) and those that did not

TABLE 3. PAIRWISE DUNN TEST RESULTS FOR PHISHING CTR AMONG OUR 11 GROUPS (POST-HOC).

	Control	FA	FA-TD2	FA-TD3	FA-TD4	FA-TD5	TD0	TD2	TD3	TD4	TD5
FA	**7.695*****										
FA-TD2	**8.885*****	1.218									
FA-TD3	**8.644*****	0.928	-0.296								
FA-TD4	**7.685*****	0.037	-1.173	-0.885							
FA-TD5	**8.411*****	0.787	-0.422	-0.130	0.745						
TD0	**4.840*****	-2.870	**-4.079*****	**-3.808****	-2.889	**-3.631****					
TD2	**6.837*****	-0.820	-2.029	-1.746	-0.852	-1.600	2.034				
TD3	**5.960*****	-1.707	-2.914	-2.636	-1.733	-2.475	1.149	-0.881			
TD4	**7.306*****	-0.393	-1.610	-1.322	-0.428	-1.177	2.478	0.430	1.000		
TD5	**7.337*****	-0.298	-1.506	-1.220	-0.334	-1.077	2.548	0.517	1.396	0.091	

* $p<.05$; ** $p<.01$; *** $p<.001$

(average phish CTR: 50.5%). This finding suggests that with a time delay people were less likely to click on phishing links. However, the effect size of time delay on phishing CTR is less than the effect of focused attention.

Looking at the data from the 960 participants whose warnings used a time delay (i.e., 2, 3, 4, or 5 seconds), a Kruskal-Wallis test did not detect any significant differences in phishing click through rate among different lengths of time delay (n.s.). This finding suggests that the length of time delay does not affect phishing click-through rate as much as having a time delay at all. Thus, hypothesis H2a is partially confirmed: *Warnings using time delays have significantly lower phishing CTR that those without time delay; there is no significant effect for the length of time delay.*

Phishing CTR lower with focused attention (H3a). Comparing all participants who saw a focused attention warning against those whose warning did not use focused attention, a Mann-Whitney U-test found significant differences in phishing click-through rates ($r=-0.1184$, $p<.001$) between participants whose warnings used focused attention (mean phish CTR: 38.5%) and warnings that did not (phish CTR: 49.2%). This finding suggests that participants were less likely to click on phishing links if their warning employed focused attention and confirms H3a: *Warnings that use focused attention have significantly lower phishing CTRs than warnings without focused attention.*

No significant differences in intrusiveness (H2b & H3b). We next examined differences in self-reported Likert scale responses for warning intrusiveness. This helps us contextualize "warning effectiveness" by incorporating participants' subjective experience with our phishing warnings. The median score for Intrusiveness across all groups was between 2 and 3, suggesting participants did not find our designs overly intrusive (see Figure 8). A Mann-Whitney U-test found no significant differences in self-reported warning intrusiveness between participants who did and did not see warnings using focused attention ($r=-0.02234$, n.s.) which confirmed H3b: *Warnings that use focused attention do not have different self-reported Intrusiveness levels than those that use focused attention.*

Surprisingly, a Mann-Whitney U-test found no significant differences in self-reported warning intrusiveness between warnings that did and did not use a time delay ($r=-0.04904$, n.s.). A Kruskal-Wallis test then also did not surface significant differences in intrusiveness between different time delays ($n.s.$). Contrary to our hypothesis, we must reject H2b: *Warnings using longer time delays do not have significantly higher values of self-reported intrusiveness.*

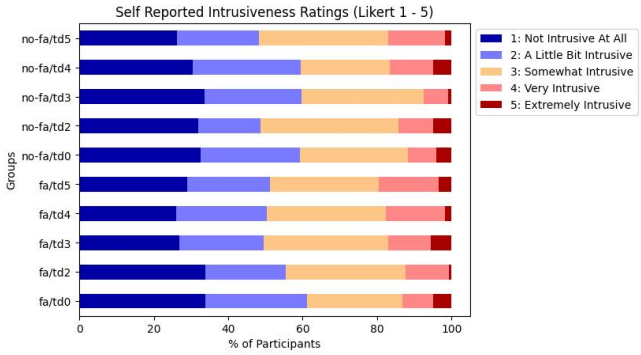

Figure 8. A proportional horizontal boxplot of self-reported Intrusiveness ratings for each treatment group.

Combined, these findings suggest that even though focused attention and time delays restrict link interactions, these measures are not perceived as overly intrusive according to self-reported Likert scale ratings.

5.2.2. Exploratory regression analyses. To better contextualize our findings regarding our pre-registered hypotheses, we followed our comparison tests with a series of exploratory regression analyses to better understand how time delay and focused attention shaped participant experience. This included mixed-effect logistic regression to identify predictors of clicking on phishing links and false positives, as well as linear regressions for the total amount of time hovered over benign and false positive links. Our regression models contained fixed effects for our link restrictions (i.e., focused attention and time delay as binary variables, and time delay values as ordinal variables), participant SeBIS scores, the three phishing URL types (i.e. gibberish, keyword-related, and brand-related domains) and demographic variables (e.g., age, occupation, etc.). We in-

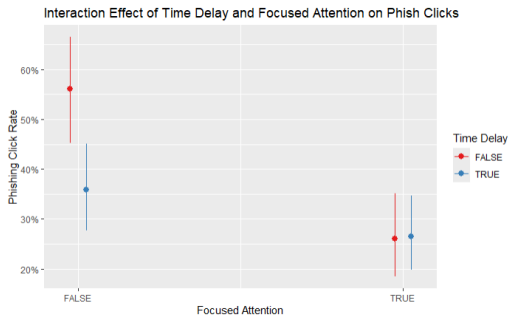

Figure 9. Interaction effect plot of focused attention and time delay on phishing click through rates.

cluded two additional fixed effects to account for different confounding variables: overall benign CTR as a measure of each participant's overall engagement with links, and the number of warnings already seen when a participant engaged with a phishing link. We use these to identify effects of warning fatigue or priming (i.e., did participants click more or less often as they encountered subsequent warnings). Random effects were the individual emails and individual hyperlinks within each email.

Focused attention and time delay most effective combined. In our first model, we examined all phishing links from participants who saw a warning (i.e., treatment group participants). We found significant main effects on phishing link CTR from both focused attention ($\chi^2(1,\ N$=3171)$=-1.29,\ p<.001$) and time delay ($\chi^2(1,\ N=3171)=-0.83,\ p<.001$). This suggests that, while both focused attention and time delay effectively dissuaded participants from clicking on phishing links, focused attention was more effective than time delay. There was also a significant interaction effect from the combination of focused attention and time delay ($\chi^2(1,\ N=3171)=0.86,\ p<.001$, see Figure 9). Warnings that used both focused attention and time delay had an average phish CTR of 37.5%. We also see in Table 3 that the baseline warning (time delay: 0; focused attention: no; phish CTR: 42.1%) has significantly higher phish CTR than all conditions that use both focused attention and time delay except for the 4-second time delay warning. Taken together, this suggests that time delay and focused attention are more effective at preventing phishing link clicks when combined.

Our regression model identified several other significant fixed effects on phishing link clicks. There was a significant effect from a participant's overall click-through rate on benign links ($\chi^2(1,\ N=3171)=3.62,\ p<.001$, mean hover time over each benign link: 1903ms), which suggests participants who were already likely to click on benign links were more likely to click on phishing links. The number of warnings seen when a participant interacted with a phishing link also had a significant effect on phishing clicks ($\chi^2(1,\ 3171)=-0.24,\ p<.001$) which suggests participants were less likely to click on phishing links the more warnings they saw.

Of our three phishing URL types, only phishing links that used the gibberish phishing URL type had a significant effect on phishing clicks ($\chi^2(1,\ 3171)=-0.28,\ p<.001$, CTR: 41.0%). This suggests that participants were less likely to click on gibberish phishing domains, but had more mixed responses to the brand-related (phishing CTR: 44.5%) and keyword-related (phishing CTR 46.1%) phishing URLs.

We also included results from the SeBIS subscales: only Proactive Awareness had a significant effect on phishing clicks ($\chi^2(1,\ 3171)=0.17,\ p<.05$). This suggests that the Proactive Awareness subscale is perhaps more relevant to identifying and avoiding phishing attacks than the Device Securement and Updating subscales.

Several of our demographic fixed effects also had significant effects on phishing clicks. Men were significantly more likely to click on phishing links ($\chi^2(1,3171)=0.39,\ p<.001$), adding to the decidedly mixed literature on the effect of gender on identifying phishing ([9], [57], [58]). Age also had a significant effect on phishing clicks ($\chi^2(1,3171)=-0.29,\ p<.001$), which suggests that older participants were also less likely to click on phishing links, though this may also be sample dependent [9].

Longer time delays potentially more effective. Next, we sought to identify the effects of particular values of time delay (i.e., 2, 3, 4 and 5 seconds). We conducted a similar regression analysis as before but filtered out participants who did not see a time delayed warning (i.e., time delay of 0 seconds). We also changed the time delay fixed effect from a binary variable to a factor to identify differences between levels of time delay. Our model identified significant main effects of focused attention ($\chi^2(1,2796)=-0.49,\ p<.001$) and a 5 second time delay ($\chi^2(1,\ 2796)=-0.14,\ p<.05$, phish CTR: 40.8%), and significant interaction effects between focused attention and 4-second ($\chi^2(1,\ 2796)=0.52,\ p<.05$, phish CTR: 41.7%) and 5-second time delays ($\chi^2(1,\ 2796)=0.59,\ p<.05$, phish CTR: 37.6%), suggesting that longer time delays prevent participants from clicking on phish.

Our regression model for time delay values also reproduced similar results to our model for focused attention and time delay. Participants were less likely to click on gibberish phishing URLs ($\chi^2(1,2796)=-0.27,\ p<.05$), if they saw multiple warnings ($\chi^2(1,\ 2796)=-0.25,\ p<.001$), if they scored higher on the SeBIS Proactive Awareness subscale ($\chi^2(1,2796)=-0.1,\ p<.05$), or if they were older ($\chi^2(1,2796)=-0.32,\ p<.001$). Conversely, participants were more likely to click on phishing links if they had high benign click-through rates ($\chi^2(1,2796)=3.5,\ p<.001$) or if they were men ($\chi^2(1,2796)=0.3856,\ p<.001$).

Time delays increase hover time. Finally, we conducted a linear regression on the total amount of time hovered over each phishing link. Our results identified main effects of time delay on hover time ($b=0.22,\ p<.01$, 2,066ms per phish link with time delay; 1,794ms without time delay) but not for focused attention ($b=0.17$, n.s., hover time with focused attention: 2,073ms per phish link; without focused attention:

1,948ms), suggesting that participants were more likely to hover over a phishing link for a longer period of time if their warning used a time delay. Our results also identified that participants were less likely to hover for a longer period of time if they had seen multiple warnings (b=−0.55, p<.001), but were more likely to hover if they scored high on the SeBIS proactive awareness scale (b=0.06, p<.05).

5.2.3. False Positives. We further examined the effect of predictors on clicking on false positive warnings.

Lower FP click rates with link restrictions. In our regression model for false positive click rates, we included focused attention and time delay as binary fixed effects. We excluded the number of warnings seen as a fixed effect since all but 30 of our participants saw the false positive as their fourth and final warning. Our regression results identified significant main effects of both focused attention ($\chi^2(1, 1187)$=−2.32, p<.001, FP CTR with focused attention: 32.8%) and time delay ($\chi^2(1, 1187)$=−0.52, p<.05, FP CTR with time delay: 41.4%), as well as significant interaction effects of time delay and focused attention ($\chi^2(1, 1187)$=1.80, p<.001, FP CTR: 36.5%). This suggests that participants were less likely to click through false positive warnings to benign websites if their warning used either or both focused attention and time delay. Consistent with our previous results, participants were more likely to click on false positive links if they had high benign link click-through rates ($\chi^2(1, 1187)$=3.97, p<.001) and were less likely to click if they were older ($\chi^2(1, 1187)$=−0.27, p<.001).

Treating time delay duration value as a factor (2, 3, 4, and 5 seconds), our regression analysis identified significant main effects of focused attention ($\chi^2(1, 949)$=−0.52, p<.05) but no significant main effect from any values of time delay nor interaction effects between focused attention and levels of time delay. This suggests that focused attention may have led to fewer participants clicking the false positive link. Otherwise, only age had a significant effect on false positive clicks ($\chi^2(1, 949)$=−0.20, p<.05), suggesting older participants were less likely to click on links with a warning.

A mixed-effect linear regression on the amount of time participants spent hovering on the false positive link found main effects of both focused attention (b=0.24, p<.05, mean FP hover time with focused attention: 4,662ms; without focused attention: 4,753ms) and time delay (b=0.3, p<.001, FP hover time with time delay: 4,907ms; without time delay: 3,914ms), and an interaction effect of time delay and focused attention (b=−0.32, p<.05, FP hover time: 4,652ms) on total false positive link hover time. This suggests that participants hovered over false positive links longer for both focused attention and time delay, and even slightly longer with time delay. Again, age also had a significant effect (b=0.06, p<.05)—older participants appear to be more deliberate when assessing hyperlinks during our study.

5.3. Effects of Phishing URL Types

Looking at phishing CTR for the three phishing URL types across groups (see Figure 10), we see similar patterns among the three URL types. The control group has higher phishing CTR than treatment groups and focused attention groups had lower phishing CTRs than non-focused attention groups. However, the median CTR values for the non-focused attention groups are higher for keyword- and brand-related URLs than for gibberish URLs.

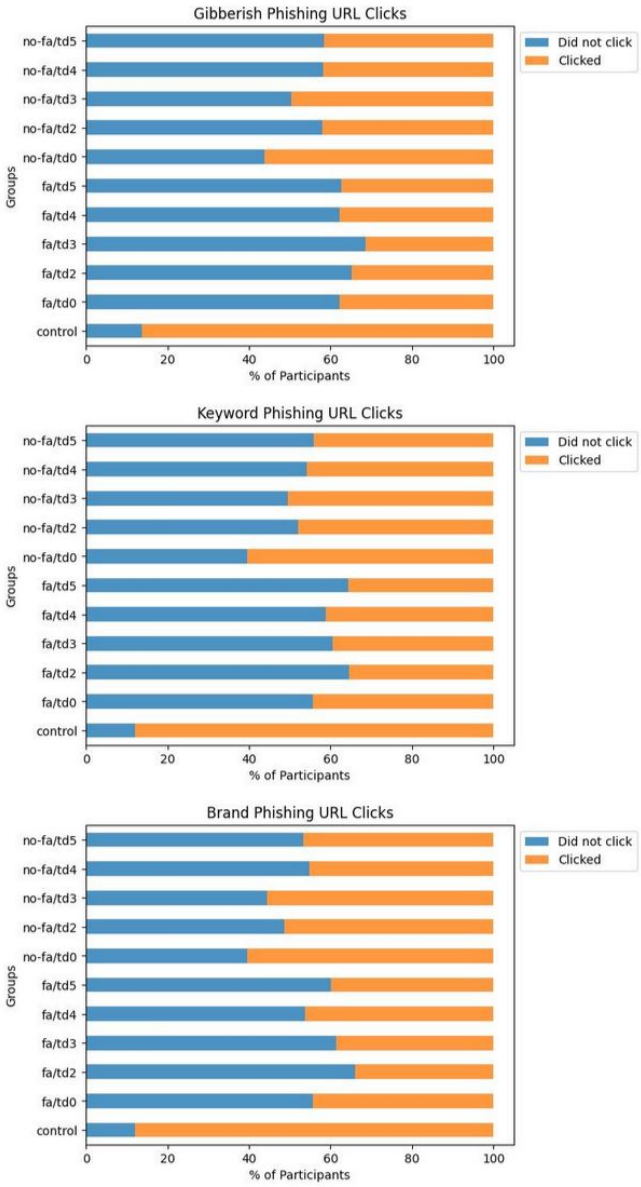

Figure 10. Phishing CTR proportional bar charts for each of the three phishing URL types (gibberish, keyword, brand) for each of our 11 groups.

We analyzed phishing click variance for each of the three phishing URLs.

Link restrictions effective against all three URL types. Mann-Whitney U-tests showed significant differences in CTR on gibberish URLs between focused attention and non-focused attention (mean gibberish CTR: 35.8% vs.

45.2%, $r=-0.1065$, $p<.001$) and between time delay and no time delay (gibberish CTR: 39.5% vs. 47.0%, $r=-0.0607$, $p<.05$). Results were similar for keyword-related phishing URLs with significant differences for focused attention versus no focused attention (mean keyword CTR: 39.2% vs. 49.7%, $r=-0.1057$, $p<.001$), and time delay versus no time delay (keyword CTR: 42.5% vs. 52.3%, $r=-0.07884$, $p<.01$). These findings also hold for brand-related phishing URLs: focused attention versus no focused attention (brand CTR: 40.6% vs. 51.8%, $r=-0.1122$, $p<.001$) and time delay versus no time delay (brand CTR: 44.6% vs. 52.2%, $r=-0.06182$, $p<.05$). This indicates that both focused attention and time delay were effective at preventing participants from clicking on all three types of phishing URLs.

5.4. Qualitative Insights

Next, we report qualitative insights on how participants assessed phishing links, usability aspects of the phishing warnings, and educational effects of the study.

5.4.1. Link assessment strategies. Participants used various methods to evaluate suspicious links. Some were expert-recommended best practices: comparing the domain name to the legitimate domain name or hovering over a link to check the status bar. Participants also reported other assessment practices, such as assessing the trustworthiness of the email or the organization, comparing the suspicious link's domain to other links in the email, checking whether the domains from the link or the email sender are familiar, or checking that a URL used HTTPS. A few participants mentioned that embedded link data within our phishing URLs made them suspicious or that emails from financial services made participants particularly cautious. Others mentioned assessment methods that required clicking on the suspicious link, such as evaluating the URL in the browser's address bar, examining the suspicious page, or trusting their existing assemblage of security processes to warn them of and keep them out of danger. P4750's response is emblematic of the problem people face when trying to identify and assess phishing links: *"'westernunion-pay.com' seems suspicious to me but for all I know, that's their actual website."* These experience-based strategies [5] could expose individuals to malware on a malicious website.

In line with our prior work [3], curiosity also played a role in how participants assessed and interacted with suspicious links. While our warning used a minimal design for experimental consistency, the lack of information about why a warning was presented made some participants curious about where the links would take them. For instance, P2960 said *"i thought [the link] was suspicious at first ... but then I got brave and clicked on the link."* These findings suggest a need to provide better information about what about a website or URL triggered display of a warning as a way to deter people from visiting malicious websites.

Participants also discussed reasons for not clicking on suspicious links. Many mentioned not wanting to put their computer or information at risk. Over 50 participants mentioned their experience with previous (incorrect) warnings. In line with prior work [5], our participants engaged in a mix of expert-recommended best practices and more contextual experience-based practices to assess suspicious links.

5.4.2. Warning usability. A few participants mentioned usability issues. Two participants believed the time delay indicated the warning was checking the link's authenticity, and once it was active that the link was safe to click on. Three participants believed the warning blocked all mouse interactions with a link.

Participants mentioned the time delay most frequently as intrusive or as something to be removed to improve the warning. This is not to say that participants did not appreciate the time delay; many participants also noted that warnings should be intrusive to some extent or that the time delay gave participants an opportunity to pause and check a link before clicking it. But given that no participants mentioned focused attention (i.e., having to click inside the warning to visit a suspicious link) as intrusive or needing to change, we take this as a small amount of evidence that time delay might be a bit more intrusive than focused attention.

5.4.3. Educational impacts. Several participants mentioned learning something through our study. Participants mentioned learning about phishing attacks and ways to identify phishing attacks. A few participants reflected on how they clicked on phishing links in the study and what that means for their browsing habits, including with online studies. These responses highlight benefits of educating participants about cybersecurity practices during and after research studies both as a social good and for reciprocity with participants, particularly when deception is used [41], [42].

6. Discussion

Our findings show how different types of link restriction can shape how people interact with phishing links and warnings. While these specific results may not translate exactly outside of our simulated inbox experience, our findings nevertheless suggest that link restrictions are an important lever to consider when developing organizational anti-phishing systems. We discuss how our findings might guide the refinement of contemporary anti-phishing warnings both in email clients (such as Proton's "Link confirmation" dialogue that appears after a link click [59]) and more broadly (e.g., Microsoft Defender's Safe Link warnings [60]).

Focused attention and time delay: effective independently and together. Our results affirm findings from prior work that both focused attention [3] and time delay [4] link restrictions reduce phishing CTR, with the additional contribution of comparing their effect on CTR independently and when combined. Participants in both focused attention and time delay warning groups clicked on phishing links significantly less than those who did not see a warning (control). Focused attention appears to be more effective than time

delays on their own, and the duration of a time delay had a negligible effect compared to having a time delay at all. The significant interaction effect between focused attention and time delay and the differences between the baseline warning (time delay: 0; focused attention: no) and the other warnings that used both time delay and focused attention suggests that combining these link restrictions further reduces phishing CTR. At the same time, we see that both link restrictions also reduced false positive CTR which suggests the effect of link restrictions on CTR is somewhat independent of whether or not the link is a phish.

Our findings have implications for contemporary anti-phishing systems. First, it is important to acknowledge that participant responses to focused attention and time delay were qualitatively and quantitatively different. Overall, it appears that different link restrictions have differing effects on hyperlink CTR, which suggests warnings might be improved by strategically matching link-restriction strategies to risk assessment levels in current contextual warnings. For instance, Mozilla Thunderbird (see bottom of Figure 11) currently utilizes link restriction by displaying a focused attention warning (which uncovers the destination URL) for links that may be phish. Thunderbird uses at least three different indicators to detect phish: obscured IP addresses using hexadecimal or octal formats, mismatched link text and host names, and emails that contain non-address book HTML forms [61]. But in terms of link restrictions, Thunderbird's warnings treat these indicators equivalently by using the same link restriction (focused attention) for all three indicators. It may be appropriate to add a time delay to the existing focused attention warning (reducing CTR and increasing friction and intrusiveness) for stronger phish indicators like an obscured IP address. This insight extends to enterprise anti-phishing solutions as well, such as Microsoft's Safe Link system [60]. For instance, it may make sense to add a time delay to Safe Link's Suspicious Warning Message (prompting readers to revisit the suspicious email) or using focused attention to couple the currently-separated hyperlink and uncovered URL in the Malicious Website warning. More research is needed on matching individual and combined link restrictions in phishing warnings with the risk levels associated with particular phishing indicators.

Link restrictions as personalized security measures. Our results also suggest the need for more personalization in information security systems, echoing calls by other researchers [16], [26], [62], [63], [64]. Time delay was noticeably more intrusive to participants than focused attention in their open responses, especially when they thought a link was legitimate (whether the link was actually legitimate or not). At the same time, some participants appreciated the time delay and being prompted to look closely at the URL. In the context of personalizing security warnings, these findings suggest that tailoring interventions to people's preferences and abilities can improve efficacy, compliance, and usability with security warnings [62].

To personalize link restrictions in phishing warnings, we can look at Proton Mail's Link Confirmation warning [59]

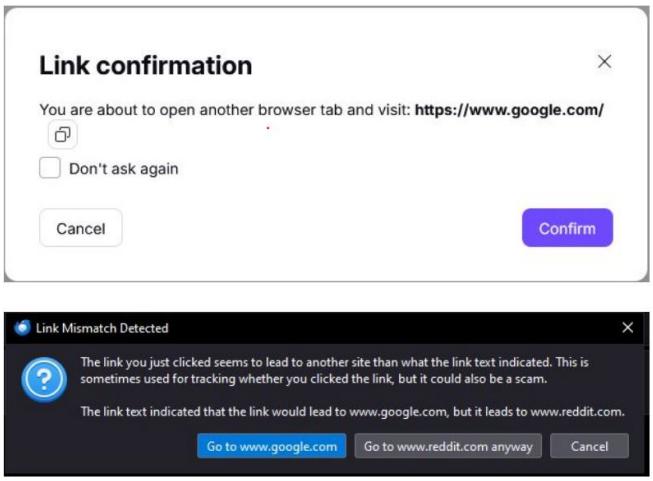

Figure 11. *Top:* Proton Mail's current security dialog. This warning uses a simple link restriction, where the dialog appears anytime a link is clicked. This is accompanied by a "Don't ask again" button, which permanently dismisses this warning for all future link clicks.
Bottom: Mozilla Thunderbird's Link Mismatch warning, which uses focused attention to draw attention to a hyperlink's true destination URL when a mismatch between the displayed hyperlink text (an anchor tag's inner HTML value) and the true destination (an anchor tag's href value).

as an example of the risks of one-size-fits-all approaches to security warnings, as well as foundations for building more personalized security experiences. By default, Proton Mail shows the Link Confirmation dialog after an email reader clicks any link and contains the true destination URL (See top of Figure 11). Proton Mail presents this security warning as an all-or-nothing approach; email readers can entertain the security dialog after every link click (even links they are sure are benign beforehand), or they can click the "Don't ask again" button and permanently dismiss the warning. This demonstrates the risk of applying link restrictions evenly across all links, as many users are indirectly funneled into permanently dismissing the warning to save time. Email readers must change their Settings (a single check box) to re-enable the warning, which they are unlikely to find unless they are explicitly looking for it.

It is here in the settings that we can imagine combining rulesets for phish identification (as Thunderbird does, described above) with different warning features (such as link restriction or placement) to enable email readers to choose both the context under which a warning appears as well as its presentation. Proton makes for a particularly useful foundation here for allowing email readers the opportunity to customize their email security in their account settings. By contrast, Thunderbird anti-phishing settings are only available through its esoteric Config Editor which requires knowledge about Thunderbird that not all email readers have. Similarly, Microsoft Defender's Safe Link settings are only available to administrators [60]. Our results show some promise in allowing individual email readers to tailor some of their user-facing security settings to their own contexts and expertise rather than trying to enact a single uniform

solution. This is neither an easy nor straightforward task, and requires more research into what types of options should be made available, and how to best onboard people (particularly those with little technical or security expertise) or what most suitable default settings might be.

7. Future Work

As discussed in our Limitations section, our experiment focused on desktop or laptop users. However, a growing portion of phishing attacks are aimed at mobile devices, such as phishing through SMS text messages (smishing) [20], [21], [65], [66]. The interaction for investigating URLs is also different on touchscreen devices than for laptops or desktops, where a long press on a hyperlink opens up a separate context menu instead of a hover. Touchscreens are much more prevalent than laptops or desktops, both globally [67] and particularly in low-income communities [68]. Even four of our participants noted that our warnings would not work for mobile devices, e.g., P5332 said: "The big problem is when you check your email on an iPhone, you cannot see the URL. They need to fix that." All of these factors point to a need for more work in anti-phishing warnings for touchscreen and mobile devices.

Accessibility is another aspect of anti-phishing warnings that requires more research. Warnings are visual cues which are not helpful for people with visual impairments. Prior work has shown how existing anti-phishing technologies pose barriers to people with visual impairments [69], [70]. More research is needed into making phishing detection technologies accessible to screen readers.

8. Conclusion

We conducted an online between-subjects experiment (n=1,320) to assess and compare the effectiveness of two link restriction methods for phishing warnings: focused attention and time delay. Our findings demonstrate that both approaches significantly reduce phishing click-through rates. Focused attention is slightly more effective than time delay, and we also see strong interaction effects when combing both link restriction types. We find no significant differences regarding the length of time delays. Our qualitative data suggests that time delay might be seen as slightly more intrusive than focused attention but we did not observe significant differences in self-reported intrusiveness ratings. Our findings demonstrate that link restrictions constitute an important lever in phishing detection systems and that whether a warning uses focused attention, time delay, or their combination could be tailored to phishing risk, operational context, and personal preferences and expertise.

Acknowledgments

The authors thank their participants, as well as the reviewers for their constructive feedback. This work has been supported by a Google Faculty Research Award and this work was partially funded by the Topic Engineering Secure Systems, subtopic 46.23.01 Methods for Engineering Secure Systems, of the Helmholtz Association (HGF) and supported by KASTEL Security Research Labs, Karlsruhe.

References

[1] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri, "Bridging the Gap in Computer Security Warnings: A Mental Model Approach," *IEEE Security & Privacy*, vol. 9, no. 2, pp. 18–26, Mar. 2011.

[2] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter, "Your attention please: Designing security-decision UIs to make genuine risks harder to ignore," in *Proceedings of the Ninth Symposium on Usable Privacy and Security*, ser. SOUPS '13. New York, NY, USA: Association for Computing Machinery, Jul. 2013, pp. 1–12.

[3] J. Petelka, Y. Zou, and F. Schaub, "Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–15.

[4] M. Volkamer, K. Renaud, and B. Reinheimer, "TORPEDO: TOoltippoweRed Phishing Email DetectiOn," in *ICT Systems Security and Privacy Protection*, ser. IFIP Advances in Information and Communication Technology, J.-H. Hoepman and S. Katzenbeisser, Eds. Cham: Springer International Publishing, 2016, pp. 161–175.

[5] R. Wash, N. Nthala, and E. Rader, "Knowledge and Capabilities that Non-Expert Users Bring to Phishing Detection," in *Seventeenth Symposium on Usable Privacy and Security ({SOUPS} 2021)*, 2021. [Online]. Available: https://www.usenix.org/conference/soups2021/presentation/wash

[6] S. Egelman, L. F. Cranor, and J. Hong, "You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings," in *CHI*. ACM, 2008, pp. 1065–1074.

[7] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. Cranor, "Crying Wolf: An Empirical Study of SSL Warning Effectiveness." *USENIX security symposium*, pp. 399–416, 2009.

[8] K. Althobaiti, K. Vaniea, and S. Zheng, "Faheem: Explaining URLs to people using a Slack bot," *Digital Behaviour Intervention for Cyber Security*, 2018.

[9] S. Baki and R. M. Verma, "Sixteen years of phishing user studies: What have we learned?" *IEEE Transactions on Dependable and Secure Computing*, vol. 20, pp. 1200–1212, 2021.

[10] D. Lain, K. Kostiainen, and S. Čapkun, "Phishing in Organizations: Findings from a Large-Scale and Long-Term Study," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 842–859.

[11] D. Akhawe and A. P. Felt, "Alice in Warningland: A {Large-Scale} Field Study of Browser Security Warning Effectiveness," pp. 257–272. [Online]. Available: https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/akhawe

[12] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter, "Your attention please: Designing security-decision uis to make genuine risks harder to ignore," in *Symposium on Usable Privacy and Security (SOUPS)*, 2013, pp. 1–12.

[13] J. Blythe, J. Camp, and V. Garg, "Targeted risk communication for computer security," in *Proceedings of the 16th International Conference on Intelligent User Interfaces*, ser. IUI '11. Association for Computing Machinery, pp. 295–298.

[14] A. Alnajim and M. Munro, "An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection," in *2009 Sixth International Conference on Information Technology: New Generations*, pp. 405–410.

[15] F. Quinkert, M. Degeling, J. Blythe, and T. Holz, "Be the Phisher – Understanding Users' Perception of Malicious Domains," in *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, ser. ASIA CCS '20. Association for Computing Machinery, pp. 263–276.

[16] A. Franz, V. Zimmermann, G. Albrecht, K. Hartwig, C. Reuter, A. Benlian, and J. Vogt, "{SoK}: Still Plenty of Phish in the Sea — A Taxonomy of {User-Oriented} Phishing Interventions and Avenues for Future Research," pp. 339–358. [Online]. Available: https://www.usenix.org/conference/soups2021/presentation/franz

[17] M. Volkamer, M. A. Sasse, and F. Boehm, "Analysing Simulated Phishing Campaigns for Staff," in *Computer Security*, I. Boureanu, C. C. Drăgan, M. Manulis, T. Giannetsos, C. Dadoyan, P. Gouvas, R. A. Hallman, S. Li, V. Chang, F. Pallas, J. Pohle, and A. Sasse, Eds. Springer International Publishing, pp. 312–328.

[18] A. Felt, R. Reeder, H. Almuhimedi, and S. Consolvo, "Experimenting at scale with google chrome's ssl warning," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[19] M. Volkamer, K. Renaud, B. Reinheimer, and A. Kunz, "User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn," *Computers & Security*, vol. 71, pp. 100–113, 2017, user-Study Between Subject Design.

[20] M. L. Rahman, D. Timko, H. Wali, and A. Neupane, "Users Really Do Respond To Smishing," in *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, ser. CODASPY '23. Association for Computing Machinery, pp. 49–60.

[21] M. Liu, Y. Zhang, B. Liu, Z. Li, H. Duan, and D. Sun, "Detecting and Characterizing SMS Spearphishing Attacks," in *Proceedings of the 37th Annual Computer Security Applications Conference*, ser. ACSAC '21. Association for Computing Machinery, pp. 930–943.

[22] Anti Phishing Working Group (APWG), "Phishing Activity Trends Reports Q2 2024." [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q2_2024.pdf

[23] C. David Hylender, P. Langlois, A. Pinto, and S. Widup, "2024 Verizon Data Breach Investigations Report."

[24] M. Mossano, O. Kulyk, B. M. Berens, E. M. Häußler, and M. Volkamer, "Influence of URL Formatting on Users' Phishing URL Detection," in *Proceedings of the 2023 European Symposium on Usable Security*, ser. EuroUSEC '23. Association for Computing Machinery, pp. 318–333.

[25] S. Patil, R. Schlegel, A. Kapadia, and A. J. Lee, "Reflection or action? how feedback and control affect location sharing decisions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. Association for Computing Machinery, pp. 101–110.

[26] Y. Wang, P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget, and N. Sadeh, "A field trial of privacy nudges for facebook," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. Association for Computing Machinery, pp. 2367–2376.

[27] J. Nicholson, L. Coventry, and P. Briggs, "Can we fight social engineering attacks by social means? Assessing social salience as a means to improve phish detection," *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[28] M. Steves, K. Greene, and M. Theofanos, "A Phish Scale: Rating Human Phishing Message Detection Difficulty," in *Workshop on usable security (USEC)*, 2019.

[29] M. L. Jensen, M. Dinger, R. T. Wright, and J. B. Thatcher, "Training to Mitigate Phishing Attacks Using Mindfulness Techniques," *Journal of Management Information Systems*, vol. 34, no. 2, pp. 597–626, 2017.

[30] M. Butavicius, K. Parsons, M. Pattinson, and A. McCormac, "Breaching the human firewall: Social engineering in phishing and spear-phishing emails," *Australasian Conference on Information Systems*, 2016.

[31] Y. Zeng, T. Zang, Y. Zhang, X. Chen, and Y. Wang, "A Comprehensive Measurement Study of Domain-Squatting Abuse," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pp. 1–6.

[32] E. Lin, S. Greenberg, E. Trotter, D. Ma, and J. Aycock, "Does domain highlighting help people identify phishing sites?" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. Association for Computing Machinery, pp. 2075–2084.

[33] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in *Proceedings of the 2007 ACM Workshop on Recurring Malcode*, ser. WORM '07. Association for Computing Machinery, pp. 1–8.

[34] M. Volkamer, K. Renaud, G. Canova, B. Reinheimer, and K. Braun, "Design and Field Evaluation of PassSec: Raising and Sustaining Web Surfer Risk Awareness," in *Trust and Trustworthy Computing*, ser. Lecture Notes in Computer Science, M. Conti, M. Schunter, and I. Askoxylakis, Eds. Springer International Publishing, pp. 104–122.

[35] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Teaching Johnny not to fall for phish," *ACM Transactions on Internet Technology (TOIT)*, vol. 10, no. 2, p. 7, 2010.

[36] G. Canova, M. Volkamer, C. Bergmann, and B. Reinheimer, "NoPhish App Evaluation: Lab and Retention Study," *Workshop on Usable Security*, 2015.

[37] S. Stockhardt, B. Reinheimer, M. Volkamer, P. Mayer, A. Kunz, P. Rack, and D. Lehmann, "Teaching Phishing-Security: Which Way is Best?" *International Conference on ICT Systems Security and Privacy Protection (IFIP SEC)*, 2016.

[38] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. Cranor, J. Hong, and E. Nunge, "Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish," *Symposium on Usable Privacy and Security (SOUPS)*, 2007.

[39] V. Braun and V. Clarke, "Using thematic analysis in psychology," in *Qualitative Research in Psychology*. Taylor & Francis Group. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa

[40] S. Egelman and E. Peer, "Scaling the security wall: Developing a security behavior intentions scale (sebis)," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 2873–2882.

[41] P. Finn and M. Jakobsson, "Designing ethical phishing experiments," in *IEEE Technology and Society Magazine*, vol. 26, no. 1, 2007, pp. 46–58.

[42] D. B. Resnik and P. R. Finn, "Ethics and Phishing Experiments," in *Science and Engineering Ethics*, vol. 24, no. 4, 2018, pp. 1241–1252.

[43] B. E. Hilbig and I. Thielmann, "On the (Mis)Use of Deception in Web-Based Research: Challenges and Recommendations," vol. 229, no. 4, pp. 225–229.

[44] T. Pfeiffer, H. Theuerling, and M. Kauer, "Click Me If You Can!" in *Human Aspects of Information Security, Privacy, and Trust*, L. Marinos and I. Askoxylakis, Eds. Springer, pp. 155–166.

[45] T. Stojnic, D. Vatsalan, and N. A. G. Arachchilage, "Phishing email strategies: Understanding cybercriminals' strategies of crafting phishing emails," vol. 4, no. 5, p. e165.

[46] J. Hong, "The state of phishing attacks," vol. 55, no. 1, pp. 74–81.

[47] S. Albakry, K. Vaniea, and M. K. Wolters, "What is this URL's Destination? Empirical Evaluation of Users' URL Reading," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 1–12.

[48] K. Althobaiti, N. Meng, and K. Vaniea, "I Don't Need an Expert! Making URL Phishing Features Human Comprehensible," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, no. 695, pp. 1–17. [Online]. Available: http://doi.org/10.1145/3411764.3445574

[49] J. Reynolds, D. Kumar, Z. Ma, R. Subramanian, M. Wu, M. Shelton, J. Mason, E. Stark, and M. Bailey, "Measuring Identity Confusion with Uniform Resource Locators," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. Association for Computing Machinery, pp. 1–12.

[50] J. Spaulding, D. Nyang, and A. Mohaisen, "Understanding the effectiveness of typosquatting techniques," in *Proceedings of the fifth ACM/IEEE Workshop on Hot Topics in Web Systems and Technologies*, 2017, pp. 1–8.

[51] PhishTank, "Phishtank: Join the fight against phishing," 2024, accessed: 2024-06-06. [Online]. Available: https://phishtank.org/

[52] B. M. Berens, F. Schaub, M. Mossano, and M. Volkamer, "Better together: The interplay between a phishing awareness video and a link-centric phishing support tool," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–60.

[53] A. J. Kimmel, "Deception in research." in *APA Handbook of Ethics in Psychology, Vol 2: Practice, Teaching, and Research.*, S. J. Knapp, M. C. Gottlieb, M. M. Handelsman, and L. D. VandeCreek, Eds. American Psychological Association, pp. 401–421.

[54] V. Distler, M. Fassl, H. Habib, K. Krombholz, G. Lenzini, C. Lallemand, L. F. Cranor, and V. Koenig, "A Systematic Literature Review of Empirical Methods and Risk Representation in Usable Privacy and Security Research," vol. 28, no. 6, pp. 43:1–43:50.

[55] G. Guzman and M. Kollar, "Income in the United States: 2022," US Department of Commerce, 2023. [Online]. Available: https://www.census.gov/content/dam/Census/library/publications/2023/demo/p60-279.pdf

[56] US Census Bureau. (2023) Income, Poverty and Health Insurance Coverage in the United States: 2022. US Department of Commerce. [Online]. Available: https://www.census.gov/newsroom/press-releases/2023/income-poverty-health-insurance-coverage.html

[57] J. G. Mohebzada, A. El Zarka, A. H. BHojani, and A. Darwish, "Phishing in a university community: Two large scale phishing experiments," in *International conference on innovations in information technology (IIT)*. IEEE, 2012, pp. 249–254.

[58] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2010, pp. 373–382.

[59] Proton Mail. (n.d.) Link confirmation. Proton Foundation. [Online]. Available: https://proton.me/support/link-confirmation

[60] C. Davis. Safe Links overview for Microsoft Defender for Office 365. Microsoft Learn. [Online]. Available: https://learn.microsoft.com/en-us/defender-office-365/safe-links-about

[61] Mozilla Foundation. Mozilla Thunderbird's Github Repository. [Online]. Available: https://github.com/mozilla/releases-comm-central/tree/master#thunderbird

[62] S. Egelman and E. Peer, "The Myth of the Average User: Improving Privacy and Security Systems through Individualization," in *Proceedings of the 2015 New Security Paradigms Workshop*, ser. NSPW '15. Association for Computing Machinery, pp. 16–28.

[63] E. Peer, S. Egelman, M. Harbach, N. Malkin, A. Mathur, and A. Frik. Nudge Me Right: Personalizing Online Security Nudges to People's Decision-Making Styles.

[64] A. Sasse, "Scaring and Bullying People into Security Won't Work," *IEEE Security & Privacy*, vol. 13, no. 3, pp. 80–83.

[65] M. Jakobsson, "Two-factor inauthentication – the rise in SMS phishing attacks," vol. 2018, no. 6, pp. 6–8.

[66] S. Mishra and D. Soni, "SMS Phishing and Mitigation Approaches," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pp. 1–5.

[67] J. Srinivasan, S. Bailur, E. Schoemaker, and S. Seshagiri, "The Poverty of Privacy: Understanding Privacy Trade-Offs From Identity Infrastructure Users in India," vol. 12, no. 0, p. 20. [Online]. Available: https://ijoc.org/index.php/ijoc/article/view/7046

[68] M. Madden, M. Gilman, K. Levy, and A. Marwick, "Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans," vol. 95, no. 1, pp. 53–126. [Online]. Available: https://heinonline.org/HOL/P?h=hein.journals/walq95&i=59

[69] G. Sonowal, K. S. Kuppusamy, and A. Kumar, "Usability evaluation of active anti-phishing browser extensions for persons with visual impairments," in *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 1–6.

[70] M. Bohlender, R. Morisco, M. Mossano, T. Schwarz, and M. Volkamer, "SMILE4VIP: Intervention to Support Visually Impaired Users Against Phishing," in *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pp. 650–657.

Appendix A.
Meta-Review

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

A.1. Summary

This paper presents a pre-registered online study on the effectiveness of security warnings in preventing users from clicking on phishing URLs. The study employed a between-subjects design, where participants were randomly assigned to one of eleven groups, one of which was a control group, and asked to interact with a virtual inbox environment, similar to Gmail, containing emails with varying types of hyperlinks (phishing, false positive, and benign). Participants were instructed to assess the functionality of each hyperlink and label the entire email as "Trash" if they found a hyperlink with an issue. The study's methodology aimed to simulate real-life scenarios and gather insights on how participants assessed phishing links, usability aspects of the warnings, and educational effects. The warning text pre-study (n=485) found that the adjective "dangerous" best described the harms of phishing and that the word "link" best described a suspicious email web address. The main study (n=1,320) results indicated that focused attention and time delay effectively prevented users from clicking on phishing URLs, but focused attention was more effective than time delay. Qualitative responses from the participants showed that time delay was more intrusive than focused attention.

A.2. Scientific Contributions

- Independent Confirmation of Important Results with Limited Prior Research
- Provides a Valuable Step Forward in an Established Field

A.3. Reasons for Acceptance

Phishing emails continue to be a commonly used and effective method of attack in many situations. This paper contributes to the knowledge of user studies on phishing link-click prevention. It is the first study to compare two previously explored defense mechanisms. The study's different hypotheses and methodology were pre-registered in OSF registries before the study began. The methodology used in the pre-survey and main study was rigorous and well-designed. The study's findings confirm that well-designed phishing warnings can prevent users from falling victim to phishing attempts. Furthermore, by collecting qualitative responses from the participants, the paper offers insights about link assessment strategies users adopt and the usability factor of phishing warnings.