

Design and Evaluation of an Anti-Phishing Artifact Based on Useful Transparency*

Christopher Beckmann, Benjamin Berens, Niklas Kühl, Peter Mayer, Mattia Mossano, Melanie Volkamer

Karlsruhe Institute of Technology, Kaiserstrasse 12, 76131 Karlsruhe, Germany
`name.surname@kit.com`

Abstract. Background: Various security interventions to support users in detecting phishing emails exist including providing the URL in a tooltip or the statusbar.

Aim: Designing and evaluating an anti-phishing artifact based on the Useful Transparency theory.

Method: We used the design science research approach for the entire process. As evaluation we ran a between-subjects study with 109 participants from the UK to determine the anti-phishing artifact effectiveness to support users distinguishing between phishing and legitimate emails.

Results: Our results show that, when compared against the state of the art security interventions (displaying the URL in the statusbar), our anti-phishing artifact increase the detection significantly, i.e. phishing detection increased from 50% to 72%.

Conclusion: Albeit further studies are required, the evaluation demonstrate that the Useful Transparency theory can result in promising security interventions. Thus, it might be worth considering it for other security interventions, too.

Keywords: Anti-phishing · Tool evaluation · Design Science Research.

1 Introduction

The Federal Bureau of Investigation (FBI) ranked phishing as 2020's most common cybercrime (see [14]) and the International Business Machines Corporation (IBM) rated it as the second most expensive cause of breaches (see [22]). Although phishing detection tools have improved over the years, users still find phishing attempts in their inbox and they will continue to do so in future. The main reasons are that (1) phishers keep developing their attack strategies and (2) legitimate messages sometimes contain phishing indicators, e.g., call to urgency.

Advanced phishing emails containing links can only be reliably detected with careful analysis of the URL behind the links. If the URL behind a link in an Amazon email is, e.g., <https://www-amazon.com> or <https://www.arnazon.jp>, the email is clearly a phishing email. However, Wash [61] showed that most

* Supported by funding from topic 46.23.01 Methods for Engineering Secure Systems of the Helmholtz Association and by KASTEL Security Research Labs.

people are not aware of this defense and Albakry et al. [1] showed that lay users have difficulties reading URLs correctly, being rarely aware that they should mainly consider the domain and top-level-domain (TLD) of a URL (e.g., for `https://www.amazon.com.host-shop.com`, “host-shop.com”).

Security experts have released and discussed various approaches to support users, i.e., technical means to detect phishing (e.g., in [64]), phishing awareness and training (e.g., in [4, 43, 49]) and tooltips containing the URL behind a link and appearing when hovering the mouse over a link (e.g., in [41, 57]), as well as exploring the reasons for phishing attacks success (e.g., in [51, 60, 63]).

We focus on designing and evaluating a novel anti-phishing artifact, which underlying idea is based on the Useful Transparency theory from Hosseini et al. [21]: Enhancing the transparency of URLs behind links in emails. We integrate so called “transparent-strings” in all emails, in most cases consisting of domain and TLD of the URL behind each link (the exceptions are explained later). Note, these strings are the only links in the emails. Figure 1 depicts an example.

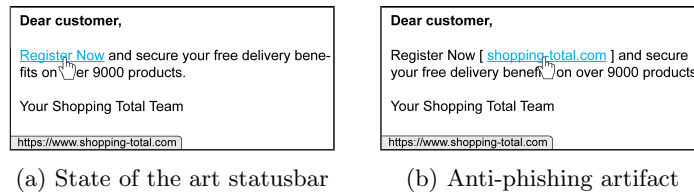


Fig. 1: Link in email, without and with our anti-phishing artifact.

Compared to the anti-phishing tooltip ideas, there are various advantages: (1) the relevant information to detect phishing URLs is displayed immediately, not only when hovering the link with the mouse. (2) We do not present the entire URL as a link, but only the transparent-string. Thereby, we thwart various URL obfuscation techniques such as subdomain-as-domain and path-posing. (3) It can be applied on both the server-side and the client-side. (4) It can be applied in the mobile context. (5) It may support visually impaired users in phishing detection, as the transparent-string is more easy to read aloud than longer URLs.

In this work, we evaluate our anti-phishing artifact in a between-subjects study with 109 Clickworker participants from the UK. We asked participants to distinguish between 28 screenshots of both phishing and legitimate emails. The study group saw the screenshots with the transparent-strings, while the control group saw unmodified email screenshots. Our results show that participants in the study group have an overall phishing recognition rate of 79.4%, against the control group’s 60.57%. Thus, applying the Useful Transparency theory (see Vossing et al. [59]), our anti-phishing artifact results in significantly better phishing detection than the baseline. Further, the control group results confirm previous research results, i.e., that most people are not aware that they should check the URL in the statusbar and have problems reading URLs cor-

rectly. We discuss how to improve the effectiveness for such attacks as directions for future work.

2 Research Design

Our overall research design is Design Science Research (DSR), as it allows to consider the theoretical and practical tasks necessary when designing and implementing IT artifacts (see [31]) and has proven to be an important and legitimate paradigm in IS research (see [19]). Acknowledging that different methodologies for design science exist (see [20, 40, 48, 52]), in the style of Kühl et al. [25] we favor a clear differentiation between an abstract “suggestion” and a concrete, more programming-specific “development”. Following Kuechler & Vaishnavi [24] a DSR project should cover: Awareness of problem (Section 3), suggestion (Section 4), development (Section 5) and evaluation (Section 6).

Our design process is informed by the kernel theory of *Useful Transparency*, from Hosseini et al. [21], who define “useful transparency” as the ability of users “to make decisions based on the provided information and act upon them” (p. 258). The theory and its relation to our approach are explored in Section 4.

3 Awareness of the Problem

Phishing definitions focus on two aspects: (1) phishing deceiving victims to click a link and share sensitive information (e.g., passwords, personal data, bank details) through authentic-looking phishing messages (e.g., [23, 28, 41, 45]) and (2) phishing spreading malware through links/attachments (e.g., [7, 15, 26, 62]). We accept both (1) and (2) as valid, but focus our work on links contained in emails.

Phishing is not a new phenomenon, but is far from being solved or under control. Numbers and damage have rather increased than decreased: FBI [14] ranked phishing as 2020’s most common cybercrime and IBM [22] as the second most expensive cause of breaches. Verizon [53] reports that 43% of all data breaches involve phishing and 95% of phishing is delivered via email. The Anti-Phishing Working Group [3] reports that the average wire-transfer loss from business email compromise in Q2 2021 is \$106,000, up from \$75,000 in Q4 2020.

As people still find phishing in their inbox, they remain one important piece of the anti-phishing measures. Simple phishing emails can be identified by checking the sender address or the plausibility of the content. Advanced attacks, however, are sent from spoofed email addresses and contain plausible content. This because the content is either obtained by re-using a legitimate emails or it is based on credible information collected, e.g., from webpages and/or social networks.

Thus, the only reliable indicator to recognize a phishing email is the URL behind the links, as shown in Garera et al. [16] and Ma et al. [30]. However, the URL is only displayed once the link is hovered with the mouse. In many desktop contexts the URL behind a link is not displayed where the users’ focus is, but in the statusbar in the lower-left corner of the browser window. Often, phishing emails also use arbitrary link text as a means to disguise the real URL, e.g., by

showing a seemingly correct URL as link text. Very advanced phishing emails may have clickable elements (e.g., form and formaction elements) that do not show any URL in the statusbar, visible only in the email source code. Phishers also use short URL and redirect services to hide the final destination, hiding it even in the source code and requiring to reach the final destination without actually opening the webpage.

Researchers have shown in Wash [61] that users are not aware of the need to check the URL behind a link. When they do, they have difficulties judging them, e.g., as shown in Albakry et al. [1]. Thus, to decrease phishing risks, email receivers need to be further supported. We do so with our anti-phishing artifact, that simplify the decision making on whether a link is safe to click or directs to a phishing page. Based on the Useful Transparency theory from Hosseini et al. [21], the anti-phishing artifact provides the information needed to judge a link in the email text, without users' actions or reading the source code.

4 Suggestion

In this section we introduce a short overview of our artifact. A full description of its working is presented in Mossano et al. [37].

We apply the Useful Transparency theory, from Hosseini et al. [21] to increase the *effectiveness* of email receivers in distinguishing between legitimate and phishing emails. We do so by enhancing the transparency of the relevant information, showing it in easy-to-read and easy-to-judge text-based links with the *transparent-string* available whenever an email is opened, without further users' actions. Thus, the relevant (and only the relevant) information is provided just-in-place, i.e., where users' focus is just before clicking a link.

The transparent-string is in most cases the domain and TLD of the original URL behind the link or, for short URLs and redirect URLs, of the final destination URL. Depending on the URL, the transparent-string can also be an IP address or, for cloud service URLs, include some subdomains to indicate that the corresponding account owner is in charge of the content, not the organization in the domain (e.g., docs.google.com). The transparent-string only provides the minimum information required to decide on the URL legitimacy. Note, the statusbar is left untouched and it shows the entire URL on mouse hover. An example of an email modified by our anti-phishing artifact is in Figure 1.

Our design has two main reasons: no extra user action is needed to get the relevant information to judge a link, as the indicator (transparent-string) is in the email text, i.e., just-in-place, as recommended in Petelka et al. [41]. Besides, the transparent-string reduces the amount of wrong decisions, as phishers can no longer trick users with subdomain attacks or the path attacks (e.g., "amazon.com.host749.com" becomes "host749.com"). Fairly, if a webpage legitimate URL is unknown or if the difference between the legitimate and phishing URL is minimal (e.g., "shop-total.com" instead of "shopping-total.com"), seeing the transparent-string would not help much. However, this would be true also without our artifact, which in turn helps those users that know the legitimate URL.

5 Development

The proposed anti-phishing artifact can either be applied centrally or locally. We decided to implement it locally as an extension for either a web browser or an email client that modifies the email right before displaying it. Although mobile email client apps could be extended too to apply our anti-phishing artifact before displaying an email, the focus of our current research is on the desktop context.

Clickable elements in emails. Various HTML elements can create links in emails: anchor-elements, form-elements, formaction-elements and area elements. From the users’ point of view, there is no difference between form-elements and formaction-elements. Note, “link” usually indicates only anchor elements, but we use it for all four types for simplicity. JavaScript could also be used to create links, but our artifact does not address it as the common email clients and web mail services block such elements. We call *link-types* how links appear to users, i.e., images, URL-like (e.g., “www.amazon.com” or “facebook.com”), and text (e.g., “Click here”). Each one could also look like a button (see Figure 2).



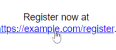

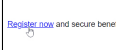


Anchor, Form, Formaction							Area Map
Image Type		URL-Like		Misc			
Generic	Button-like	Generic	Button-like	Generic	Button-like		
Artifact	<i>Unmodified</i>						
							
	Image Link: [example.org]	Image Link: [example.org]	Register now at [example.com]	[example.org]	Register now [example.com] and secure benefits.?	Start Now: [example.org]	Area Link: [example1.com] Area Link: [example2.com]

Fig. 2: Before and after applying our anti-phishing artifact

Abstract algorithm. Our anti-phishing artifact deals with all the different elements and link-types. First it resolves the transparent-string for all URLs in an email. Then, it proceeds based on how the link is integrated (see Figure 2). As HTML and CSS are relatively rich languages, emails can be very complex and we cannot rule out that the anti-phishing artifact makes them unreadable. Hence, we implemented a toggle function to undo all substitutions on demand.

Resolving the transparent-string Using short URL services or a redirect services the original URL is not the final destination. Thus, we first check whether such services were used. If so, we apply the functionality proposed in Volkamer et al. [57] to reveal the final destination URL without loading the corresponding web page. Afterwards, the final destination URL is treated as the original URL.

Next, we extract the host from the URL with the functionality proposed in Volkamer et al. [57] and check whether it is an IP address. If yes, the IP address is displayed as the transparent-string. If not, the transparent-string is the domain and the TLD of the URL, using the Mozilla Foundation’s Public Suffix list to do so (see Mozilla [38]), as proposed in Volkamer et al. [57]. Last, the transparent-string is checked against potential homographic attacks, handling non-ASCII characters by replacing them in the transparent-string with so-called puny code. Note, this is the approach adopted by programs such as Google Chrome 51+.

Specific attack strategies If our approach is adopted, phishers may adapt their strategies to it. Hence, we asked several security researchers to think of potential attacks. They proposed what we call the *doctored-pruned-URL*: phishers could try to confuse users by putting the link only on parts of a URL-like text, e.g., “amazon.com”. This link would be modified to “amazon.com [book-657.jp]”.

6 Evaluation

6.1 Methodology

Our main goal is to evaluate the usability of our proposal in the private context. With respect to *effectiveness* in distinguishing between legitimate and phishing emails, our anti-phishing artifact is based on the Useful Transparency theory. Correspondingly, we want to confirm the following hypothesis:

H-effective. Our anti-phishing artifact helps participants to significantly *better* distinguish between legitimate and phishing emails than without.

Furthermore, we investigate efficiency and satisfaction by answering the following research questions:

RQ-efficient. How efficient are participants with and without our anti-phishing artifact in distinguishing between legitimate and phishing emails?

RQ-satisfaction. How do participants rate our anti-phishing artifact on the System Usability Scale (SUS) compared to the statusbar?

Study Design. We designed a between-subjects study with two groups: one *study group* (SG) and one *control group* (CG). The SG saw emails with our artifact, i.e., the complete URLs are displayed in the statusbar and the transparent-string is added to the email text. The CG saw unmodified email, i.e., the complete URLs are displayed in the statusbar and the email text is unmodified.

We used a *role-play approach*: participants were asked to answer according to a specific scenario (details in paragraph “Scenario”). We combined this role-play with a *quiz-like approach*: participants saw email screenshots and were asked to

distinguish whether it was a phishing email or not. A binary choice is representative of private users’ real world judgment conditions. There are questionnaires to measure security awareness, e.g., Vishwanath et al. [54]. Yet, they are not focused on phishing and target awareness, rather than decision making in realistic situations. Hence, we believe that a quiz-like approach is more appropriate for our research goal. The participants saw static screenshots, with the cursor hovering over the link and displaying the browser’s statusbar with the URL behind the link (as in Reinheimer et al. [43], see Section 6.1). We are aware that this makes security the participants’ primary, if not only, task. Yet, this allowed us to evaluate various phishing attacks without running simulated phishing campaigns, avoiding the challenges shown in Volkamer et al. [58] and Pirocca et al. [42]. It also allowed us to run an online study, reaching a higher number of participants and avoiding issues with the COVID-19 restrictions on lab studies.

The emails are in the Chrome web browser, in MS Windows 10. This is the most common combination of desktop operating system and browser in the UK, according to Statcounter [46,47]. The emails are seen in the Gmail web interface.

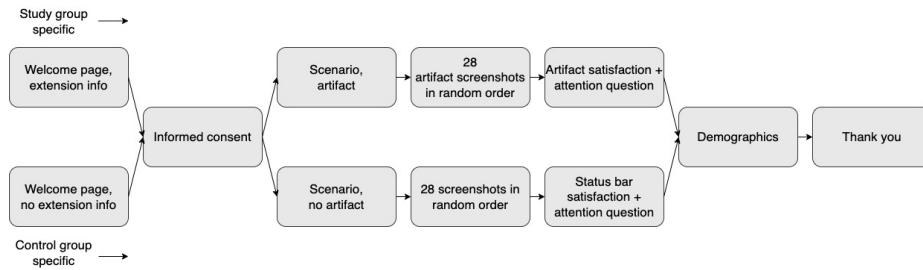


Fig. 3: Study procedure.

The *study procedure* is depicted in Figure 3. The different steps are:

Welcome. The participants recruitment through SoSci Survey is detailed in Section 6.1. After clicking the SoSci Survey link, all participants saw one of two welcome pages. We explained the evaluation and what their task was. The SG read that we were evaluating a new browser extension¹ and how it helps distinguishing between legitimate and phishing emails; no explanation about the extension working was given. We told the CG that we were interested in determining their skill to distinguish between legitimate and phishing emails. Potential limitations caused by the security focus are addressed in Section 7.

Informed Consent. We asked the participants to not use external materials or help to limit external influences. We also explained them that a low performance

¹ We mentioned a browser extension to avoid confusion for the slightly different email appearance. We don’t believe this biased the participants, as their primary task was distinguishing between phishing and legitimate emails.

had no negative consequence. We also informed the participants of the attention questions that, if answered incorrectly, would have barred them from continuing and be compensated. We added that they could have terminated the study at any time, without providing any reason, by closing the browser window. If this was the case, we would not have used their data. Lastly, we informed the participants that their data would have been processed and stored in Germany, not the UK.

Scenario. We decided to use a role-play approach, slightly different depending on the group. We told both groups to distinguish between phishing and legitimate emails as if they were a fictional persona named “David T. Jones”, recently moved to the UK. Relevant information about David was provided. We again informed the SG that David installed a new browser extension to help him recognize phishing emails. No explanation about the extension working was given.

Judging Screenshots. We asked the participants to distinguish among 28 email screenshots, presented in random order, the phishing from the legitimate ones. Every email was for a service activation. The SG saw email screenshots with the transparent-links, while the CG saw the same images without modifications. More information on the emails and their selection is provided in Section 6.1.

Satisfaction. Each group saw a page with one of the screenshots from the study with either the statusbar or our the transparent-string being highlighted. We then asked participants to answer the System Usability Scale (SUS)². The participants were asked to answer each item as themselves on a 5-points Likert scale, ranging from strongly disagree to strongly agree. Note, we added one additional attention question asking participants to select a specific item of the Likert scale.

Demographics. We asked our participants various question about themselves: Highest study degree and its field, which email provider they mainly use, which software they usually check emails with, whether they received any anti-phishing training or informed themselves on how to check detect malicious emails, what type of anti-phishing training / awareness material they used to learn, and how long ago did they receive the training or informed themselves last.

Thank you. We thanked them for their participation and clarified that we modified some of the legitimate emails to better fit the study, so they should not have considered the screenshots as perfect examples of official emails. We also provided our contacts again, in case they were needed.

Email Selection. We decided to show the same amount of phishing and legitimate emails. We exclusively included advanced phishing emails, i.e., phishing emails that can only be identified as such using the URL behind the link. This

² The SUS is a common tool used to evaluate the usability of systems and products, initially developed by Brooke [8]. We used the SUS version from Bangor et al. [5]

means that we selected legitimate emails and changed the URL behind the relevant link to get a phishing email. We also modified the link type to equally cover the different UI cases from Section 5.

The following *three dimensions* are relevant for us to decide which phishing emails to include³: the different UI cases, the URL obfuscation technique, and the sending organization. As it can be seen in Figure 2, there are seven different *situations from a user interface perspective*. However, the area-element one is related to the image-generic case and is only very rarely used in the email context. Therefore, we decided to focus on the remaining six situations (two per link-type – see Figure 2) for the user study. Furthermore, we distinguish four *URL obfuscation techniques*⁴: arbitrary-URL (i.e., the URL domain is an arbitrary name or IP), subdomain-as-domain (i.e., the host name is placed in the subdomain part of the URL), path-posing (i.e., the host name is placed in the path of the URL), and typo-swapping (i.e., similar looking domain but, e.g., spelled with letters in different positions or similar looking characters, e.g., rn instead of m). Note, we decided to use URLs with HTTPS protocol for both phishing and legitimate email screenshots. This was done for three reasons: Firstly, nowadays most phishing websites use SSL/TLS, as reported in APWG [3]. Secondly, participants may judge URLs legitimacy on their protocol alone, as reported in Alsharnouby et al. [2]. For the same reasons, HTTPS-only was used in Albakry et al. [1], Volkamer et al. [57] and Peteleka et al. [41]. Ultimately, as shown in Oest et al. [39], almost 86% of successful phishing attacks use HTTPS.

Thus, we have *12 different phishing cases* to be considered (4 URL obfuscation techniques x 3 link-types). We decided to use different sending organizations for each of the 12, i.e., 12 different legitimate emails that were changed into phishing ones. Considering that the UK has a specific double format top level domain (.co.uk), we decided to include one such legitimate URL per obfuscation technique, i.e., ending with such top level domain. These 12 organizations were identified based on the top ALEXA UK pages. In addition to the four URL obfuscation techniques, we also considered *two additional attack cases*: mismatch attack and the doctored-pruned-URL attack. The mismatch attack was also studied in Chiew et al. [12] and Caputo et al. [11], and it was described as a link showing a URL address different than the one behind it. Ideally, all URL obfuscation techniques in Table 1 should be studied in combination with these two attack cases. However, the resulting number of screenshots would be too high. Therefore, we decided to use the arbitrary URL obfuscation technique in combination with these two attack cases. For these two additional cases we use two new sending organizations, again identified based on the top ALEXA page for the UK. Thus, in total 14 phishing emails are studied. See Table 1 for a description of each of the *28 emails*.

³ Note, for the user study, we do not consider form and form-action elements. This has two reasons: CG would have no chance to decide about phishing or not for those emails as the URL would not appear in the statusbar. Furthermore, for the UI/screenshots it does not make a difference which of the two elements is used

⁴ These are similar techniques to those studied in, e.g., [29, 44, 55, 56].

Obfuscation technique	Link type	Organization	Legitimate URL	Phishing URL
Arbitrary URL	Image - Generic	BBC	https://www.bbc.co.uk	https://www.linkyzt.com
	URL-like - Generic	Netflix	https://www.netflix.com/browse?lnktrk=EMP&g=4F4D261316D39C280880331...	https://www.host745.com/browse?lnktrk=EMP&g=4F4D261316D39C280880331...
	Misc - Button-like	Spotify	https://wl.spotify.com/	https://129.13.152.9
Subdomain-as-Domain	Image - Button-like	Google	https://accounts.google.co.uk/signin/v2/identifier?service=accountsettings...	https://accounts.google.co.uk.nimsky57.ru/identifier?service=accountsettings...
	URL-like - Button-like	Facebook	https://www.facebook.com/	https://www.facebook.com.host547.com/
	Misc - Generic	Instagram	https://www.instagram.com/activate	https://www.instagram.com.3nk317rc.com/activate
Path-Posing	Image - Generic	Ebay	https://rover.ebay.co.uk/rover/2	https://www.mpls.com/www.ebay.co.uk
	URL-like - Generic	Wikipedia	https://en.wikipedia.org/wiki/Special:ConfirmEmail/a784d79322cb80d4f1127...	https://www.host875.com/en.wikipedia.org/wiki/Special:ConfirmEmail/a784d79322...
	Misc - Button-like	Zoom	https://us05web.zoom.us/activate?code=xTk7ww9F_p-zq4eTrrNExMcEGiD...	https://www.providershop58.com/us05web.zoom.us/activate?code=xTk7ww9F_p...
Typo-Swapping	Image - Button-like	Amazon	https://www.amazon.co.uk/	https://www.amzaon.co.uk/
	URL-like - Button-like	The Guardian	https://profile.theguardian.com/verify-email/q1fjo-KOUgAkzWwRpyPxS1...	https://profile.theguardain.com/verify-email/q1fjo-KOUgAkzWwRpyPxS1...
	Misc - Generic	Microsoft	https://www.microsoft.com/en-gb/activate/jjuP9kjj3uH78dhsuuy&89kiOhEyp9m	https://www.mircosoft.com/en-gb/activate/jjuP9kjj3uH78dhsuuy&89kiOhEyp9m
Mismatch	URL-like - Generic	Daily Mail	https://www.dailymail.co.uk/registration/activate.html?email=jones.t.david88%40...	https://www.jiorlikniski.cn/registration/activate.html?email=jones.t.david88...
Doctored-Pruned-URL	Misc - Generic	UK Government	https://www.gov.uk/confirm	https://www.uhszhiklo.cz

Table 1: List of URLs used in the email screenshots used in the user study

Recruitment, Data Protection, and Ethics. We recruited UK participants using the panel service “Clickworker”. According to Cohen [13], without sufficient information – as it is the case for our study – a medium effect size helps not to over- or underestimate the expected effect size. Therefore, we decided to plan for medium effect size. We assumed to use a T-test for independent groups, for the test strength analyses with G*Power. In addition to the effect size, we set the test power to 0.8 and the alpha error to 0.05. Hence, we calculated a sample size of 51 participants per group. To avoid falling below this limit due to exclusion, we set the number of participants per group to 60 to have a buffer.

Based on pre-tests, we expected the study to be finished in 30 minutes. We wanted to pay the participants based on the UK minimum wage. However, there is no unified minimum wage, rather the remuneration is based on age and role, as shown in GOV.UK [18]. Since the participant selection was random and no age groups were pre-defined (other than participants had to be 18 or above), we decided to use the latest (at the time, the one from April 2020) minimum wage

⁴ The legitimate mismatch for the CG contains a typo in the link-text dddailymail.co.uk. The legitimate doctored-pruned-URL only has “confirm” as a link.

for the oldest age group (25+): £8,72/hour. Once considered the time required to complete the study, the participants received £4,36 (8,72 : 60 = 4,36 : 30).

We used SoSci Survey to collect the data, as they are compliant with the *European Data Protection Regulation* (GDPR). However, in the UK the GDPR principles were added to the *Digital Act* of 2018, creating what is now known as UK-GDPR. We informed the participants that their data is stored and processed in Germany. We provided them with a link to the privacy policy of SoSci Survey and a contact person among the researchers. The study description was submitted for consideration and approved by the ethical board of our university. as part of the review, the data protection officer of our university checked and approved both the informed consent and the overall study design.

6.2 Results

Participants and Data Cleaning. We have 126 complete datasets (not considering the participants that were directly excluded because they failed the attention question). We performed the following data cleaning steps: we excluded three participants because they judged 100% of the screenshots as either all legitimate or all phishing. We removed one outlier in sensitivity. We removed, for the same reason, nine participants in criterion⁵. Lastly, we removed four outliers because of the time. These were removed because they excessively skewed our results by violating the 1.5 times interquartile range distance in both directions (see [50]). This left us with an overall dataset of 109 participants, 53 in the CG and 56 in the SG. The average age for the CG was 36.94 with SD 9.1, ranging from 20 to 57 years and for the SG was 36.98 with SD 11.7, ranging from 18 to 67 years. The education of the CG versus the SG was 11 versus 18 high school, 27 versus 26 with a bachelor, both ten with a master, two versus none PhD, and three versus two other. Table 2a shows the email services the participants used. Table 2b represents the web clients the participants usually use to read their emails. 18 CG participants and 24 SG ones stated that they previously participated in an anti-phishing training or informed themselves about it.

Analysis Methods.

H-effective. We employed the Signal Detection Theory (SDT) to measure our participants' skill to distinguish between phishing and legitimate emails. SDT have already been used in various studies on phishing identification (e.g., in [9, 10, 32, 33, 36, 43]). SDT uses two variables, signal (phishing emails) and noise (legitimate emails), to calculate various outputs. In line with the aforementioned researches, we decided to look for two values: sensitivity (d') and criterion (C). We defined sensitivity as the skill to successfully distinguish between phishing emails (signal) and legitimate ones (noise). The large d' , the higher the participants' skill is. We use criterion (C) to determine the participants' tendency

⁵ The signal detection theory used to measure effectiveness considers sensitivity and criterion (see Section 6.2).

Email Service	CG	SG
Gmail	43	48
Yahoo	3	4
BT	3	1
iCloud	2	0
Own server	0	1
Other	2	2

Clients	CG	SG
Chrome	34	42
Outlook	12	17
Apple Mail	8	4
Firefox	5	6
Edge	4	5
Thunderbird	2	1
Other	2	4

(a) The email provider usually used. (b) Clients usually used to read emails.

Table 2: Users email services and clients.

		CG		SG	
		Mean	SD	Mean	SD
Effectiveness	Phish	50.0%	19.4	71.7%	21.7
	Legit	71.2%	14.4	87.1%	13.1
	Overall	60.6%	12.3	79.4%	13.3
	Sensitivity	0.6	0.7	2.3	1.5
	Criterion	0.3	0.3	0.4	0.6
Efficiency		511.7 sec	241.6	533.4 sec	275.7

Table 3: Descriptive statistics for both groups.

while distinguish emails. The closer C is to 0, the less tendency in the answers in direction of either phishing or legitimate emails exists. On the one hand, the more positive C is, the more a participant was over-cautious, and identified legitimate emails as phishing one (more false positives). On the other hand, the more negative C is, the more a participant showed over-confidence, identifying more phishing emails as legitimate (false negatives).

We calculated assumptions relevant for the SDT parameters, i.e., equal variance and Gaussian distribution. For sensitivity ($F = 13.848, p < 0.001$) and criterion ($F = 11.523, p = 0.001$) the assumption of equal variances is violated and therefore we report the results for the Welch t-test. We then calculated the parameters for sensitivity and criterion per participants. Afterwards, we calculated the mean values per measurements. To evaluate *h-effective*, we used independent t-tests to check the differences among participants' sensitivity and criterion. For each independent t-test, we checked the assumptions for sensitivity and criterion.

Efficiency. We used the average time spent on all screenshots. To evaluate efficiency, we started with an exploratory analysis looking at the descriptive data. As we had no knowledge about a potential difference between the SG and CG for such a "short" evaluation, we tested the hypothesis that there might be a significant difference between both groups (without a clear trend). We calculated

assumptions relevant for the efficiency, i.e., equal variance ($F = 1.628, p = 0.205$) and Gaussian distribution. Therefore we tested two-tailed and used an independent t-test to check the differences among participants' efficiency.

Satisfaction. Regarding the SUS (described in Section 6.1), we followed the guidelines on how to score it shown in Lewis [27]. We started off with an exploratory analysis looking at the descriptive data. As we had no previous knowledge about a potential difference between the SG and CG for such a “short” evaluation, we moved on with testing the hypothesis that there might be a significant difference between both groups (without a clear trend). We calculated assumptions relevant for the satisfaction, i.e., equal variance ($F = 0.616, p = 0.434$) and Gaussian distribution. Therefore we tested two-tailed and used an independent t-test to check the differences among participants' satisfaction.

Analysis Outcome. All the descriptive results are provided in Table 3.

Effectiveness. The SG participants ($M = 2.29, SD = 1.45$) demonstrated significantly better sensitivity scores, $t(107) = 7.682, p < 0.0001$, when compared to the CG participants ($M = 0.62, SD = 0.73$). The effect size for this analysis ($d = 1.448$) was found to exceed Cohen's convention for a large effect ($d = .80$) (see [13]). The participants in the SG ($M = 0.37, SD = 0.58$), when compared to the participants in the CG ($M = 0.31, SD = 0.35$), demonstrated no significant difference for the criterion scores, $t(107) = 0.683, p = 0.491$. However, as the criterion of the SG is almost neutral, demonstrating no significant difference between both groups means we can accept *h-effective*.

We also looked at the participants' performance for the individual email screenshots. The descriptive results are provided in Table 4. The Google phishing email has the worst performance for the CG (subdomain-as-domain, 17% correct answers). In comparison, the Microsoft phishing email was the worst-performing screenshot for the SG with about 34% (typo-swapping). The Facebook phishing email showed the best performance with 77% (subdomain-as-domain) for the CG. For the SG, the best result is the Dailymail phishing screenshot with 91% (mismatch) and Facebook with 89% (subdomain-as-domain). The doctored-pruned-URL attack achieved better results in the SG (77%) than in the CG (53%). Of the legitimate email screenshots, the Dailymail one achieved by far the lowest score for the CG with 32%. For the SG, the eBay email screenshot scored the lowest with 73%. The screenshot that performed worst in the CG for legitimate examples had a strange sender email address (registration@and.co.uk for dailymail.co.uk). Note, this was not due to our modifications. Without the artifact and without checking the statusbar, this might have caused a wrong judgment, as a common recommendation is to check for strange email addresses.

Efficiency. The SG took slightly longer ($M \approx 533$ seconds) to judge all screenshots compared to the CG ($M \approx 511$ seconds). We checked for significant difference between both groups and there is none, $t(107) = 0.436, p = 0.663$.

Obfuscation technique	Link type	Organization	Control Group				Study Group			
			Legit		Phish		Legit		Phish	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
Arbitrary URL	Image - Generic	BBC	94%	23	45%	50	96%	18	71%	46
	URL-like - Generic	Netflix	81%	40	60%	49	95%	23	79%	41
	Misc - Button-like	Spotify	77%	42	36%	48	84%	37	75%	44
Subdomain-as-Domain	Image - Button-like	Google	83%	38	17%	38	91%	29	73%	45
	URL-like - Button-like	Facebook	64%	48	77%	42	79%	41	89%	31
	Misc - Generic	Instagram	62%	49	47%	50	84%	37	59%	50
Path-Posing	Image - Generic	Ebay	59%	50	60%	49	73%	45	84%	37
	URL-like - Generic	Wikipedia	62%	49	64%	48	95%	23	80%	40
	Misc - Button-like	Zoom	66%	48	55%	50	79%	41	86%	35
Typo-Swapping	Image - Button-like	Amazon	83%	38	40%	49	89%	31	59%	50
	URL-like - Button-like	The Guardian	89%	32	40%	49	96%	19	46%	50
	Misc - Generic	Microsoft	60%	49	38%	49	86%	35	34%	48
Mismatch	URL-like - Generic	Daily Mail	32%	47	68%	47	82%	39	91%	29
Doctored-Pruned-URL	Misc - Generic	UK Government	81%	40	53 %	50	91%	29	77%	43

Table 4: Percentage and standard deviation of correct judgments per example.

System Usability Scale (SUS). Based on Bangor et al. [5, 6], a SUS value above 71.4 on adjective ratings scale is considered at least “good”. The SG ($M = 72.54$) is on average slightly below the CG ($M = 77.88$). However, the artifact is sufficient usable. We checked whether there is a significant difference between both groups and there is none, $t(107) = 1.812$, $p = 0.073$.

7 Discussion

Effectiveness. Our study results show that participants with the artifact perform significantly better when distinguishing between legitimate and phishing emails containing a dangerous link. An unexpected positive outcome is the large effect size. As stated in Section 6.1, we decided for a medium effect size due to the absence of information on the novel artifact. However, we found that the distance between SG and CG ($d=1.44$ 95% CI [-1.868 to -1.022]) not only exceeded our conservative approach, but also Cohen’s large effect size of 0.8. Hence, the difference between the groups’ performances is highly significant. This is very positive, as the study design we used is not favorable to the artifact: The idea behind the Useful Transparency theory (and our the artifact) is to have the critical information integrated into the email body so that the transparent-string is visible before deciding to click on a link. However, in our study design, the users did not interact with the email screenshots by themselves, i.e., the URL was already displayed in the statusbar for the CG too, without participants moving the mouse to the link. For future work, we plan to study if the difference between the two groups increases further when participants need to interact with the email to see the URL in the statusbar.

Our study shows that the artifact supports users in particular in both mismatch cases: (1) when phishers use the mismatch obfuscation technique to trick users into clicking a trustworthy looking URL in the email body. (2) When senders accidental cause a mismatch as an honest mistake – with 68% in the CG to 91% in the SG for (1) and with 32% in the CG to 68% in the SG for (2). Our findings also confirm past research results from [2]: Some of the worst performing examples were ones with the typo-swapping obfuscation technique. While we could argue that this obfuscation technique is not very realistic, as big services such as Google or Meta are continuously searching for domains similar to the legitimate ones (see [17, 34, 35]), it is worth improving the performance rate further as future work, e.g., by displaying the transparent-string differently, e.g., “a r n a z o n . c o m”.

One results that can be inferred is that adding the relevant information just-in-place, i.e., next to the link, help users distinguishing between legitimate and phishing emails, as shown in Petelka et al. [41]. We plan to investigate further this effect as future work, e.g., checking if tooltips have similar results.

Efficiency and System Usability Scale. The results show that there is no significant difference between both groups with regard to efficiency. For our study setting this seems not to be that surprising, as the screenshots already show the URL behind a link, putting the CG in nearly the same starting position as the SG, i.e., the relevant information is available without first moving the mouse to a link. Furthermore, participants have not received any explanations about our anti-phishing artifact. Whether a transparent-link is more efficient, in particular after having received some explanations and/or after having used it for some time, is left for future work. Participants’ SUS rating of the artifact is good, if we consider that it is a novel approach and that they have not received any explanations. The perceived usability is comparable to the one of the status quo (having the URL displayed in the statusbar).

Limitations. Our study design has security as its primary goal, an unlikely situation in real life. However, considering that our focus was to test a new approach’s effectiveness, we argue that having security as the primary focus still bears useful results. The first step is to perform better in such a situation; the next one is – as soon as it is allowed due to the COVID-19 restrictions – to replicate the study in a lab environment, with a cover story and fewer phishes to make it more realistic. Furthermore, it is worth noticing, that although participants primary task was security, the CG did not perform much better than guessing (60% overall hit rate and 50% phishing detection). This confirms past research results that users lack awareness of the importance of checking the URL in the statusbar before clicking the link (see Wash [61]), and that people struggle with URLs in general (see Albakry et al. [1]). This also underlines the need for approaches like our anti-phishing artifact. Also, we acknowledge that the emails were all confirmations of new accounts. Giving the fact that phishes could only be detected when checking the URL and the transparent-string respectively. Thus, for the purpose of the study the actual content of

the email was not that important. The emails were realistic, given the scenario presented. Furthermore, such emails lack the emotional pressure that, e.g., a fake bill would deliver. Similar to Reinheimer et al. [43], we use emails twice: As a phishing email and a legitimate one. Like Reinheimer et al. [43], we also believe that the impact of this choice is limited, as they were shown in random order. We choose email from organizations in the Alexa UK top webpages to reduce the effect of unknown services – and their legitimate URLs – on the results. Admittedly, we could not be sure of the familiarity of each participant with them, as we did not interviewed them about this. A possible solution to this could have been to use organizations created *ad-hoc* for this study. We believe, although, that this would not have solve the issue, as fictitious organization would have extended the unfamiliarity to every single participant, instead of some of them. We could have add the legitimate URL above each email, so that the participants could have compared those with the one showed. However, this would have greatly diminished the difficulty of the task and it would have been completely unrealistic (i.e., no legitimate URL is shown in real-life situations). We decided to use only one link for each email to help control for interfering factors. However, in real-life it is normal to received emails with multiple links, which might cause disruption when our artifact is applied. This could be solved by the toggle function, as it would return the email to its intended, original layout. Such evaluation, however, was besides the scope of this study and it will be covered in future works.

8 Conclusion

We developed a novel anti-phishing approach based on the Useful Transparency theory from Hosseini et al. [21]. The idea is to substitute all links present in an email with what we call a “transparent-string”. Thereby, the artifacts presents the relevant information (and only that) about the link where the users’ focus is. Furthermore, it enables checking the destination of the link also in case of form-elements and formactions-elements (in which no URL is displayed in the statusbar), as well as in case of short URLs and redirect URLs (without putting users at risk). We conducted an online survey with UK participants to evaluate our approach with respect to effectiveness in distinguishing between phishing and legitimate emails, efficiency and perceived usability. Our results show that by incorporating the Useful Transparency theory into our artifact, we were able to propose an approach which supports users in significantly better distinguishing between phishing and legitimate emails. Furthermore, the artifact does not lead to a delay in decision making (and, thus, has no negative impact on the users’ performance) and does not decrease the perceived usability of emails.

References

1. Albakry, S., Vaniea, K., Wolters, M.K.: What is This URL’s Destination? Empirical Evaluation of Users’ URL Reading, p. 1–12. ACM, NY, USA (2020), <https://doi.org/10.1145/3313831.3376168>

2. Alsharnouby, M., Alaca, F., Chiasson, S.: Why phishing still works: User strategies for combating phishing attacks. *Int. J. Hum. Comput. Stud.* **82**, 69–82 (2015). <https://doi.org/10.1016/j.ijhcs.2015.05.005>
3. APWG: Phishing Activity Trends Report (2021), https://docs.apwg.org/reports/apwg_trends_report_q2_2020.pdf
4. Arachchilage, N.A., Flechais, I., Beznosov, K.: A game storyboard design for avoiding phishing attacks. In: SOUPS '14. p. 2 (2014)
5. Bangor, A., Kortum, P., Miller, J.: An empirical evaluation of the system usability scale. *Int. J. Hum.-Comput. Int.* **24**(6), 574–594 (2008). <https://doi.org/10.1080/10447310802205776>
6. Bangor, A., Kortum, P., Miller, J.: Determining what individual sus scores mean: Adding an adjective rating scale. *JUX* **4**(3), 114–123 (2009)
7. Benenson, Z., Gassmann, F., Landwirth, R.: Unpacking spear phishing susceptibility. In: FC '17 (2017). https://doi.org/10.1007/978-3-319-70278-0_39
8. Brooke, J.: SUS: a “quick and dirty” usability. *Usability Evaluation in Indust.* **189**(3), 189–194 (1996)
9. Butavicius, M.A., Parsons, K., Pattinson, M.R., McCormac, A., Calic, D., Lillie, M.: Understanding susceptibility to phishing emails: Assessing the impact of individual differences and culture. In: HAISA '17. pp. 12–23 (2017), <http://www.cscan.org/openaccess/?paperid=354>
10. Canfield, C., Fischhoff, B., Davis, A.: Using signal detection theory to measure phishing detection ability and behavior. In: SOUPS '15 (2015)
11. Caputo, D.D., Pflieger, S.L., Freeman, J.D., Johnson, M.E.: Going spear phishing: Exploring embedded training and awareness. *IEEE Secur. Priv.* **12**(1), 28–38 (2014). <https://doi.org/10.1109/MSP.2013.106>
12. Chiew, K.L., Chang, E.H., Sze, S.N., Tiong, W.K.: Utilisation of website logo for phishing detection. *Comput. Secur.* **54**, 16–26 (2015). <https://doi.org/https://doi.org/10.1016/j.cose.2015.07.006>
13. Cohen, J.: *Statistical power analysis for the behavioral sciences*. Routledge, NY, USA (2013)
14. FBI: 2020 Internet Crime Report (2021), https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf
15. Filipczuk, D., Mason, C., Snow, S.: Using a Game to Explore Notions of Responsibility for Cyber Security in Organisations. In: CHI '19. pp. 1–6 (2019). <https://doi.org/10.1145/3290607.3312846>
16. Garera, S., Provos, N., Chew, M., Rubin, A.D.: A framework for detection and measurement of phishing attacks. In: WORM '07. p. 1 (2007). <https://doi.org/10.1145/1314389.1314391>
17. Google: Report domain name abuse - Google Domains Help (2021), <https://support.google.com/domains/answer/10093434?hl=en>
18. GOV.UK: National Minimum Wage and National Living Wage rates (2020), <https://www.gov.uk/national-minimum-wage-rates>
19. Gregor, S., Hevner, A.R.: Positioning and presenting design science research for maximum impact. *MIS Q.* **37**(2), 337–356 (2013). <https://doi.org/10.25300/MISQ/2013/37.2.01>
20. Hevner, A., Chatterjee, S.: Design science research in information systems. In: DESRIST '12. pp. 9–22 (2010). https://doi.org/10.1007/978-1-4419-5653-8_2
21. Hosseini, M., Shahri, A., Phalp, K., Ali, R.: Four reference models for transparency requirements in information systems. *Requir. Eng.* **23**(2), 251–275 (2018). <https://doi.org/10.1007/s00766-017-0265-y>

22. IBM: Cost of a Data Breach Report 2021 (2021), <https://www.ibm.com/security/data-breach>
23. Kirlappos, I., Sasse, M.A.: Security education against phishing: A modest proposal for a major rethink. *IEEE Secur. Priv.* **10**(2), 24–32 (2012)
24. Kuechler, W., Vaishnavi, V.: A framework for theory development in design science research: multiple perspectives. *J AIS* **13**(6), 395 (2012). <https://doi.org/10.17705/1jais.00300>
25. Kühn, N., Mühlthaler, M., Goutier, M.: Supporting customer-oriented marketing with artificial intelligence: automatically quantifying customer needs from social media. *Electron. Mark.* **30**(2), 351–367 (2020). <https://doi.org/10.1007/s12525-019-00351-0>
26. Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L.F., Hong, J., Nunge, E.: Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System. In: *CHI '07*. pp. 905–914 (2007). <https://doi.org/10.1145/1240624.1240760>
27. Lewis, J.R.: The system usability scale: Past, present, and future. *Int. J. Hum.-Comp. Int.* **34**(7), 577–590 (2018). <https://doi.org/10.1080/10447318.2018.1455307>
28. Lin, E., Greenberg, S., Trotter, E., Ma, D., Aycocock, J.: Does domain highlighting help people identify phishing sites? In: *CHI '11*. p. 2075 (2011). <https://doi.org/10.1145/1978942.1979244>
29. Lin, E., Greenberg, S., Trotter, E., Ma, D., Aycocock, J.: Does domain highlighting help people identify phishing sites? In: *CHI '11*. p. 2075–2084 (2011). <https://doi.org/10.1145/1978942.1979244>
30. Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Beyond blacklists: learning to detect malicious web sites from suspicious urls. In: *KDD '09*. p. 1245–1254 (2009). <https://doi.org/10.1145/1557019.1557153>
31. March, S.T., Smith, G.F.: Design and Natural Science Research on Information Technology. *Decis. Support Syst.* **15**(4), 251–266 (1995). [https://doi.org/10.1016/0167-9236\(94\)00041-2](https://doi.org/10.1016/0167-9236(94)00041-2)
32. Martin, J., Dubé, C., Coovert, M.D.: Signal detection theory (sdt) is effective for modeling user behavior toward phishing and spear-phishing attacks. *Hum. Factors* **60**(8), 1179–1191 (2018). <https://doi.org/10.1177/0018720818789818>
33. Mayhorn, C.B., Nyeste, P.G.: Training users to counteract phishing. *WORK* **41 Suppl 1**, 3549–52 (2012). <https://doi.org/10.3233/wor-2012-1054-3549>
34. Meta: Protecting People from Domain Name Fraud (2020), <https://about.fb.com/news/2020/03/domain-name-lawsuit/>
35. Meta: Protecting People From Imposter Domain Names (2020), <https://about.fb.com/news/2020/06/imposter-domain-names/>
36. Moreno-Fernández, M.M., Blanco, F., Garaizar, P., Matute, H.: Fishing for phishers. improving internet users' sensitivity to visual deception cues to prevent electronic fraud. *Comput. Hum. Behav.* **69**(C), 421–436 (2017). <https://doi.org/10.1016/j.chb.2016.12.044>
37. Mossano, M., Berens, B., Heller, P., Beckmann, C., Aldag, L., Mayer, P., Volkamer, M.: SMILE - Smart eMall Link Domain Extractor. In: *SPOSE '21*. p. 403–412 (2022). https://doi.org/10.1007/978-3-030-95484-0_23
38. Mozilla Foundation: Public Suffix List (2020), <https://publicsuffix.org/>
39. Oest, A., Zhang, P., Wardman, B., Nunes, E., Burgis, J., Zand, A., Thomas, K., Doupé, A., Ahn, G.J.: Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. In: *CSS '20*. pp. 361–377 (2020)

40. Peffers, K., Rothenberger, M., Tuunanen, T., Vaezi, R.: Design science research evaluation. In: DESRIST '12. pp. 398–410 (2012). https://doi.org/10.1007/978-3-642-29863-9_29
41. Petelka, J., Zou, Y., Schaub, F.: Put your warning where your link is: Improving and evaluating email phishing warnings. In: CHI '19. p. 1–15 (2019). <https://doi.org/10.1145/3290605.3300748>
42. Pirocca, S., Allodi, L., Zannone, N.: A toolkit for security awareness training against targeted phishing. In: ICISS '20. pp. 137–159 (2020). https://doi.org/10.1007/978-3-030-65610-2_9
43. Reinheimer, B., Aldag, L., Mayer, P., Mossano, M., Duezguen, R., Lofthouse, B., Von Landesberger, T., Volkamer, M.: An investigation of phishing awareness and education over time: When and how to best remind users. In: SOUPS '20. pp. 259–284 (2020)
44. Reynolds, J., Kumar, D., Ma, Z., Subramanian, R., Wu, M., Shelton, M., Mason, J., Stark, E., Bailey, M.: Measuring identity confusion with uniform resource locators. In: CHI '20. p. 1–12 (2020). <https://doi.org/10.1145/3313831.3376298>
45. Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L.F., Hong, J., Nunge, E.: Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In: SOUPS '07. pp. 88–99 (2007)
46. Statcounter: Desktop Browser Market Share United Kingdom (2020), <https://gs.statcounter.com/browser-market-share/desktop/united-kingdom>
47. Statcounter: Desktop Operating System Market Share United Kingdom (2020), <https://gs.statcounter.com/os-market-share/desktop/united-kingdom>
48. Teixeira, J.G., Patrício, L., Tuunanen, T.: Advancing service design research with design science research. *J. Serv. Manag.* **30**(5), 577–592 (2019). <https://doi.org/10.1108/JOSM-05-2019-0131>
49. Tschakert, K.F., Ngamsuriyaroj, S.: Effectiveness of and user preferences for security awareness training methodologies. *Heliyon* **5**, e02010 (2019). <https://doi.org/10.1016/j.heliyon.2019.e02010>
50. Upton, G., Cook, I.: Understanding statistics. Oxford University Press, Oxford, UK (1996)
51. Vance, A., Jenkins, J.L., Anderson, B.B., Bjornn, D.K., Kirwan, C.B.: Tuning Out Security Warnings: A Longitudinal Examination of Habituation Through fMRI, Eye Tracking, and Field Experiments. *MIS Q.* **42**(2), 355–380 (2018). <https://doi.org/10.25300/MISQ/2018/14124>
52. Venable, J., Pries-Heje, J., Baskerville, R.: Feds: a framework for evaluation in design science research. *Eur. J. Inf. Syst.* **25**(1), 77–89 (2016). <https://doi.org/10.1057/ejis.2014.36>
53. Verizon: Data Breach Investigations Report (2021), <https://enterprise.verizon.com/resources/reports/2021-data-breach-investigations-report.pdf>
54. Vishwanath, A., Neo, L.S., Goh, P., Lee, S., Khader, M., Ong, G., Chin, J.: Cyber hygiene: The concept, its measure, and its initial tests. *Decis. Support Sys.* **128**, 113–160 (2020). <https://doi.org/https://doi.org/10.1016/j.dss.2019.113160>
55. Volkamer, M., Renaud, K., Canova, G., Reinheimer, B., Braun, K.: Design and field evaluation of passsec: Raising and sustaining web surfer risk awareness. In: TRUST '15. pp. 104–122 (2015). https://doi.org/https://doi.org/10.1007/978-3-319-22846-4_7
56. Volkamer, M., Renaud, K., Gerber, P.: Spot the phish by checking the pruned url. *Inf. Comput Secur.* **24**(4), 372–385 (2016). <https://doi.org/10.1108/ICS-07-2015-0032>

57. Volkamer, M., Renaud, K., Reinheimer, B., Kunz, A.: User experiences of torpedo: Tooltip-powered phishing email detection. *Comput. Secur.* **71**, 100–113 (2017). <https://doi.org/https://doi.org/10.1016/j.cose.2017.02.004>
58. Volkamer, M., Sasse, M.A., Boehm, F.: Analysing simulated phishing campaigns for staff. In: *Comput. Secur.* pp. 312–328 (2020). https://doi.org/https://doi.org/10.1007/978-3-030-66504-3_19
59. Vössing, M., Kühn, N., Lind, M., Satzger, G.: Designing transparency for effective human-ai collaboration. *Inf. Syst. Front.* pp. 1–19 (2022). <https://doi.org/10.1007/s10796-022-10284-3>
60. Wang, J., Li, Y., Rao, H.R.: Overconfidence in phishing email detection. *JAIS* **17**(11), 1 (2016). <https://doi.org/10.17705/1jais.00442>
61. Wash, R.: How experts detect phishing scam emails. *Proc. ACM Hum.-Comput. Interact.* **4**(CSCW2), 160:1–28 (2020). <https://doi.org/10.1145/3415231>
62. Wash, R., Cooper, M.: Who Provides Phishing Training? In: *CHI ‘18* (2018). <https://doi.org/10.1145/3173574.3174066>
63. Wright, R.T., Jensen, M.L., Thatcher, J.B., Dinger, M., Marett, K.: Research note—influence techniques in phishing attacks: An examination of vulnerability and resistance. *Inf. Syst. Res.* **25**(2), 385–400 (2014). <https://doi.org/10.1287/isre.2014.0522>
64. Zhu, E., Chen, Y., Ye, C., Li, X., Liu, F.: OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network. *IEEE Access* **7**, 73271–73284 (2019). <https://doi.org/10.1109/ACCESS.2019.2920655>

A System Usability Scale

SUS scale used
I think that I would like to use this feature/ <i>extension</i> frequently.
I found this feature/ <i>extension</i> unnecessarily complex.
I thought this feature/ <i>extension</i> was easy to use.
I think that I would need assistance to be able to use this feature/ <i>extension</i> .
I found the various functions in this feature/ <i>extension</i> were well integrated.
PLEASE, FOR THIS QUESTION, SELECT STRONGLY AGREE - 5.
I thought there was too much inconsistency in this feature/ <i>extension</i> .
I would imagine that most people would learn to use this feature/ <i>extension</i> very quickly.
I found this feature/ <i>extension</i> very awkward to use.
I felt very confident using this feature/ <i>extension</i> .
I needed to learn a lot of things before I could get going with this feature/ <i>extension</i> .

Table 5: System usability scale used with research question. The italics is what the study group saw. Capitalized is the attention question added to the scale.