Influence of URL Formatting on Users' Phishing URL Detection

MATTIA MOSSANO, Karlsruhe Institute of Technology, Germany
OKSANA KULYK, IT-University Copenhagen, Denmark
BENJAMIN M. BERENS, Karlsruhe Institute of Technology, Germany
ELENA M. HÄUSLER, Karlsruhe Institute of Technology, Germany
MELANIE VOLKAMER, Karlsruhe Institute of Technology, Germany

Despite technical advances in anti-phishing protection, in many cases the detection of phishing URLs largely depends on users manually inspecting the links found in suspicious emails. One solution proposed to support users in doing so is to use a URL formatting that focuses their attention on critical URL parts, such as domain and top-level-domain (called "who-area"). While this solution has been implemented in several software products (e.g., browsers web address bar), research on its effectiveness with regard to phishing URL detection is currently limited. To investigate the extent to which different kinds of URL formatting support users to detect phishing URLs, we conducted an online study (n = 200) using interactive email screenshots with tooltips showing two previously evaluated URL formatting (called "Who-Area Highlighting" and "Who-Area Only"). A group with unmodified URLs (called "Plain URL") acted as control. We did not find any significant difference between the URL formatting within our sample, with successful phishing URL detection rates ranging from 71% (for the unmodified URL) to 76% (for showing only the URL who-area). As we designed a study with a "best-case" scenario to detect a medium effect size (Cohen's r = .30), it is then likely that the URL formatting effect on phishing URL detection would be even smaller in the real world. Still, we found that other factors had a significant effect on users' phishing URL detection, namely, gender-related factors and forcing users' attention on the URL for 4-5 seconds. However, these results are only preliminary and need further investigation.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Security and privacy → Usability in security and privacy**.

Additional Key Words and Phrases: URL formatting, domain highlighting, anti-phishing, user support

ACM Reference Format:
Mattia Mossano, Oksana Kulyk, Benjamin M. Berens, Elena M. Häusler, and Melanie Volkamer. 2023. Influence of URL Formatting on Users' Phishing URL Detection. In . ACM, New York, NY, USA, 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Anti-Phishing Working Group [3] reports the highest number of phishing attacks since they started to collect data in 2003. Albeit Kaspersky [15] notes that new vectors are on the rise (e.g., messenger apps), Verizon [35] shows that emails are still the preferred phishing delivery method. Basic phishing emails are usually stopped by email filters, easily detectable by checking the sender address or given away by the implausibility of their content. Yet, email filters require time to adapt to new attacks, allowing recent phishing emails to reach the users' inbox. Moreover, Pandove et al. [23] show that attackers can spoof sender addresses or use hacked legitimate accounts to avoid suspicion. Additionally,

Manuscript submitted to ACM

Bhavsar et al. [5] show that attackers avoid implausible content by cloning legitimate emails. We call *advanced phishing emails* those emails capable of avoiding email filters, employing spoofing and cloning techniques, and containing links.

Usually, the most effective way to detect advanced phishing emails is to check the URL behind the links, as URLs cannot be spoofed like sender addresses. Yet, attackers have developed multiple techniques to confuse users, like creating URLs similar to legitimate ones, as shown, e.g., in Reynolds et al. [28]. Besides, there are two further challenges with checking the URL behind a link: first, Alsharnouby et al. [2] and Wash [39] show that users lack awareness of the importance of such checks. Second, several studies (e.g., [1, 2, 11, 45]) found that users lack understanding of which URL parts are critical to consider (i.e., second-level-domain and top-level-domain, hereafter *who-area*).

Petelka et al. [25] and Volkamer et al. [37, 38] already address the first challenge. They show that just-in-place and just-in-time interventions displaying URLs in tooltips allow users to focus on the URL even when lacking awareness of its importance, both in email clients and in browsers. This led to a significantly higher detection of advanced phishing emails than showing URLs in the statusbar. Yet, the second challenge, i.e., how to best support users to correctly read a URL, is still open. Thus, our goal is to investigate different *URL formatting* (i.e., how the URL is displayed) in tooltips and evaluate their effectiveness in supporting users to detect phishing URLs in advanced phishing emails.

Xiong et al. [43] and Volkamer et al. [36] address a similar question in their research. However, they focused on URL formatting in Mozilla Firefox web address bar, asking participants to decide whether a web site was phishing or not based on it. Their results suggest a pattern regarding URL formatting: making the who-area to stand out (hereafter, *Who-Area Highlighting*) is more effective than showing the plain URL and showing a pruned URL (i.e., only display the who-area; hereafter, *Who-Area Only*) is more effective than Who-Area Highlighting. We use "suggest", instead of "show", because Xiong et al. [43] and Volkamer et al. [36] have several study design differences (e.g., use of eye-tracking) and they did not considered the same URL formatting (the only common one being Who-Area Highlighting).

Thus, a new study is required to evaluate all three URL formatting in the same study set-up. To this end, we conducted a between-subjects, online user study with 200 German participants that investigated the users' phishing URL detection support of the two aforementioned URL formatting shown in a tooltip[1]. Note, phishing detection is a complex process with several steps, heavily dependent on the program used to check emails and the presence of security add-on(s). Our set-up is not complex enough to allow for an ecologically valid research of it. Instead, we isolate a specific aspect of phishing detection (i.e., URL analysis) and focused on it to investigate the effects of URL formatting.

Our results show that neither Who-Area Highlighting nor Who-Area Only have a statistically significant effect on the participants' detection of phishing URLs within our sample. Still, even if URL formatting is not significantly effective, other factors show a significant influence, i.e., gender-related factors and the time spent hovering the link. The latter, in particular, suggests that forcing users to read URLs for some seconds could be a determining factor to thwart successful phishing attacks. Yet, further investigation is required to reach conclusive results on the gender-related factors.

After the introduction, we outline related work on URL formatting in section 2. We then present our methodology in section 3 and our results in section 4. The discussion of our findings is in section 5 and we conclude in section 6.

2 RELATED WORKS

We reviewed the studies focused on URL formatting – that is, making certain parts of URLs standing out – and their effectiveness in supporting phishing URL detection, i.e., in helping users to notice that the destination web site is different than expected. In particular, we identified two types of related work in the literature:

[1]Note, albeit the study methodology is presented in English, the study was conducted in German with country specific URLs.

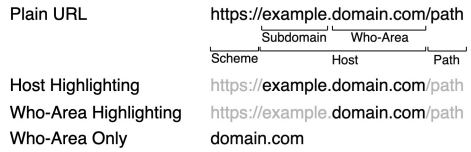

Fig. 1. Parts of a URL and examples of URL formatting from literature.

- Studies which main focus is on the evaluation of URL formatting (e.g., [16, 28, 36, 43]).
- Studies that include URL formatting as a part of a broader evaluation, e.g., as one of several design decisions in anti-phishing interventions (e.g., [22, 25, 38, 41, 42, 44]).

While studies of the second type provide useful context for the use of URL formatting, they only offer limited insights regarding different URL formatting comparative effectiveness. Hence, we focus on the studies of the first type to inform our research questions and study design, in particular, the choice of URL formatting to compare. We present both study types in the following subsections.

2.1 Studies that Focus on URL Formatting

Various studies focused specifically on studying the impact of different URL formatting on phishing URL detection (e.g., [16, 28, 36, 43]). In particular, these studies investigated three URL formatting (shown in Fig. 1): (1) *Host Highlighting*, that shows the whole URL, but everything except the host is grayed out; (2) *Who-Area Highlighting*, showing the whole URL with everything grayed out but the who-area; (3) *Who-Area Only*, which only shows the who-area.

Host Highlighting was investigated by Lin et al. [16], who conducted a within-subjects study in two phases: first, participants were presented with screenshots of sixteen web pages and asked in think-aloud sessions whether they believed them to be legitimate. Then, their attention was drawn to the web address bar and they judged the legitimacy of the same screenshots again. While the authors concluded from both the think-aloud and the exit interviews that the formatting was not increasing participants' phishing detection, no statistical analysis was conducted.

We decided against evaluating Host Highlighting because this URL formatting might be less effective against certain types of phishing tricks compared to both Who-Area Only and Who-Area Highlighting. For example, if a phishing URL masquerades as Google using the "subdomain-as-domain" technique (e.g., `google.com.megahost.ru`), Host Highlighting would highlight the whole host. Both Who-Area Only and Who-Area Highlighting, instead, would highlight only the who-area (e.g., `megahost.ru`), indicating to the user that the website does not belong to Google.

The Who-Area Highlighting and the Who-Area Only formatting were investigated by Volkamer et al. [36], adapting and expanding Lin et al.'s methodology by conducting a between-subjects comparison of the phishing detection of two groups of participants. One group saw screenshots of web sites that applied the Who-Area Highlighting formatting to the URL in the web address bar (as done in Mozilla Firefox), while the other group saw the same web sites using the Who-Area Only formatting in the web address bar. Just as in Lin et al., Volkamer et al. evaluated whether the participants were more likely to distinguish between phishing and legitimate URLs when directed towards the web address bar. Their results showed that, for both kinds of URL formatting, participants were significantly more likely to correctly detect phishing URLs after drawing attention to the web address bar. Furthermore, participants were significantly more likely to detect phishing links when their attention was drawn to the Who-Area Only formatting, rather than to the Who-Area Highlighting one. Yet, there were no significant differences in detection rates for legitimate messages, neither in the within-subjects comparison, nor the between-subjects one.

Following Volkamer et al.'s study, we evaluated the same URL formatting, adding a control group with unmodified URLs for comparison (i.e., Plain URL). Yet, we decided against explicitly pointing participants' attention towards the URL via study instructions. Instead, we decided to rely on the "just-in-time" and "just-in-place" guidelines from Petelka et al. [25] to direct the participants' attention to the URL. Thus, we present the URLs via tooltips that appear when the participants hover on a link, as opposed to the browser web address bar.

Xiong et al. [43] too extended the study from Lin et al., by conducting two experiments. In the first experiment, the participants were divided in two groups, one seeing web pages with Plain URL in the web address bar and the other with the URL using the Who-Area Highlighting formatting. They were shown six web pages (half phishing and half legitimate) in an online setting and asked to rate them on a 5-points Likert scale ranging from 1 (unsafe) to 5 (safe). In the second experiment participants' interactions with the web site screenshots were measured in a lab via an eye tracking device. Both experiments followed two phases, using two different set of six web pages. In the second phase, participants were asked to focus on the web address bar. Xiong et al. report that in both experiments, when the attention was drawn to the web address bar, the participants' phishing URL detection increased. Yet, the first experiment did not find any significant statistical difference between the detection rate of the Who-Area Highlighting group and the Plain URL group between Phase 1 and Phase 2. This was further confirmed in Phase 2 of the second experiment, which showed that, when drawing attention to the web address bar, the participants did not focus on the formatting, but rather on the presence/absence of other security indicators, such as the lock icon or HTTP/HTTPS.

Our study therefore builds upon these results, together with the results from Volkamer et al., evaluating whether users' phishing URL detection effectiveness can be improved via drawing their attention to the URL given appropriate formatting (as shown in Volkamer et al.) or, confirming the results from Xiong et al. [43], whether additional knowledge on how to read the URL is required for such improvements.

Reynolds et al. [28] adopted a different methodology than the previous studies, evaluating users' phishing URL detection in isolation, i.e., out of context. One group saw *URL Parsing Help*, i.e., applying together technical solutions (e.g., puny-codes), scheme elision (i.e., HTTP/HTTPS was hidden), and Host Highlighting. The formatting had a statistically significant effect on the phishing URL detection of the participants. Yet, as the study evaluated the effect of URL formatting alongside other solutions, the formatting effect is challenging to infer from the results. Furthermore, the formatting used by Reynolds et al. still suffers of the same issue with subdomain-as-domain tricks described before.

For comparison, we report in Table 1 the results of Phase 2 of both Volkamer et al. [36] and Xiong et al. [43].

Table 1. Related works results when attention is focused. Abbreviations: Part. = Participants, Ex. = Experiment, PL = Plain URL, WAH = Who-Area Highlighting, WAO = Who-Area Only. Note, Xiong et al. did not conduct intra-phase statistical analysis, only inter-phases.

Study		**Part.**	**Legitimate**			**Signif.**	**Phishing**			**Signif.**
			PL	*WAH*	*WAO*		*PL*	*WAH*	*WAO*	
Volkamer et al. [36]		189	–	67.45%	68.61%	No $p = 0.78$	–	49.50%	59.61%	Yes $p = 0.019$
Xiong et al. [43]	*Ex.1*	320	82.50%	78.50%	–	–	37.00%	46.20%	–	–
	Ex.2	32	87.50%	75.00%	–	–	50.00%	68.80%	–	–

2.2 Studies that Include URL Formatting

Other studies implement URL formatting in different warnings or tools, without evaluating the URL formatting effectiveness as part of their research questions. There are several examples of such works (e.g., [22, 25, 38, 41, 42, 44]). In particular, our attention was drawn to one work, as their results directly impact the effectiveness of URL formatting.

Petelka et al. [25] evaluated the effects of various link-centric warnings on phishing susceptibility in an online between-subjects study. They considered three types of warnings: a banner (similar to Gmail), a link-centric warning (tooltip) of their own design with a warning sign and color-coded threat-level (i.e., red if a link was phishing), and a browser warning appearing after clicking the link (similar to Google Chrome). The link-centric group was further divided into "with" and "without" forced attention, i.e., in the forced attention group, only the link in the tooltip was clickable. A group without any warning was their baseline. They used participants' click-through rate (i.e., how many participants reached the phishing web page) to determine phishing susceptibility. Their results showed that all warning conditions reduced participants' click-through rate, compared to the baseline. As Petelka et al. admitted as successful phishing only cases where a participant landed on the phishing web page, the most effective warning was the browser one. This means that the browser warning participants clicked on the link, reached the browser warning, but did not proceed further to the phishing web page. Yet, it might be that a browser warning would not appear if the URL is not yet block-listed. Hence, the ideal scenario is that users do not click on the link at all, as this would protect them also from those attacks that would not be blocked by a browser. This considered, the data showed that the link-centric participants had the lowest number of clicks on the link of all, in particular the forced attention group. Thus, we agree with Petelka et al. that the most promising solution is the link-centric warning, i.e., the tooltip.

Given Petelka et al.'s results, and the results on attention in Volkamer et al. [36] and Xiong et al. [43] (see section 2.1), we focused on link-centric solutions too, excluding any works that did not (namely, [22, 41, 42, 44]).

Only one study fitted our constraint: Volkamer et al. [38], who evaluated a Mozilla Thunderbird and Firefox add-on called TORPEDO. TORPEDO checks the URL of hovered links and shows the URL in a tooltip just-in-place and just-in-time, i.e., it follows Petelka et al.'s guidelines. The who-area of the URL is highlighted in bold font and further information is shown on request. This information tells users to disregard other indicators (e.g., HTTPS or sender name) and to only focus on the who-area. The "further information" pane also presents some examples of possible attacks, such as subdomain-as-domain, extension (e.g., `paypal-secure.com`) or typos (e.g., `mircosoft.de`). Finally, TORPEDO stops users from clicking for three seconds, forcing them to take time to carefully read the URL, again, following Petelka et al.'s guidelines. TORPEDO was evaluated in a between-subjects online study with two groups: one saw screenshots of emails with the URL in the status bar, and the other saw screenshots of emails showing the TORPEDO tooltip. The results show that the TORPEDO participants were statistically significantly better than the status bar group.

In our study we decided to implement forced attention as it was done in Volkamer et al. for TORPEDO, although we reduced the time to one second, instead of three. Further information on the reasons behind this choice are in section 3.5. It might also look like we included a similar URL formatting as TORPEDO, i.e., Who-Area Highlighting. Yet, TORPEDO uses bold font to highlight the who-area and it does not gray out other URL parts, hence the two URL formatting are only superficially similar. We decided to not include TORPEDO formatting for the same reason we excluded the one considered in Reynolds et al. [28] (see section 2.1): the effect on the phishing URL detection might be due to other features of TORPEDO (e.g., the forced attention), not to the URL formatting itself. Thus, based on the results of both Volkamer et al. [38] and Petelka et al. [25], plus those of Volkamer et al. [36] and Xiong et al [43], we decided to follow just-in-place and just-in-time guidelines to draw attention to the URL formatting by using tooltips.

3 METHODOLOGY

We present here our research questions, study design, ethical considerations and recruitment.

3.1 Research Questions

Our research questions are based on the results of the related work presented in section 2.

As a broad goal, we want to explore the results from related work in section 2.1 in a different context, i.e., showing the URL in a tooltip rather than in the web address bar. This is because, as specified in section 2.2, according to Petelka et al. [25] a tooltip draws users' attention to the URL formatting (contrary to a web address bar), enhancing the detection of phishing URLs. Moreover, we want to determine if hovering time has any influence on phishing URL detection. Note that phishing detection is a complex process, made of several steps and heavily context dependent. Hence, it is important to notice that our set-up is not complex enough to allow for an ecologically valid research on phishing detection as a whole. We instead isolate a specific aspect of phishing detection (i.e., URL analysis) and focus on it to investigate the effect of URL formatting. Specifically, we investigate the following two research questions:

RQ1. – **Is there a best URL formatting for tooltips among the two considered to support phishing URL detection?** The question stems from the apparent pattern of Volkamer et al. [36]'s and Xiong et al. [43]'s results, i.e., Who-Area Only allows to detect more phishing URLs than Who-Area Highlighting, and Who-Area Highlighting allows to detect more phishing URLs than Plain URL. We formulate three hypotheses:

- $H1_{WAH>PL}$: Participants seeing Who-Area Highlighting formatted tooltips detect significantly more phishing URLs than participants seeing Plain URL formatted tooltips.
- $H2_{WAO>WAH}$: Participants seeing Who-Area Only formatted tooltips detect significantly more phishing URLs than participants seeing Who-Area Highlighting formatted tooltips.
- $H3_{WAO>PL}$: Participants seeing Who-Area Only formatted tooltips detect significantly more phishing URLs than participants seeing Plain URL formatted tooltips.

RQ2. – **Does hovering time have a beneficial effect on phishing URL detection?** The question stems from the results of Volkamer et al. [38], who block participants from clicking a link for 3 seconds to force them to take time to check the URL. As no evaluation of time delay is present in their study, we want to see if hovering time has an effect on the ability to correctly detect phishing URLs, as presented in section 3.5.

3.2 Study Groups

We considered the following study groups (for examples of the specific URL formatting, see Fig. 2):

Plain URL. The tooltip shows URLs without modifications. This act as the control group of the study. The URL formatting is the standard used in status bars (e.g., Mozilla Thunderbird, Google Chrome) and most tooltips (e.g., Apple Mail in MacOS, Microsoft Outlook in Windows).

Who-Area Highlighting. The tooltip shows the entire URL, but everything except the who-area is grayed out. This URL formatting was also used in Volkamer et al. [36] and Xiong et al. [43].

Who-Area Only. The tooltip shows only the who-area to the users, without any other element. This URL formatting was also used in Volkamer et al. [36].

3.3 Survey Design

We designed a between-subjects online user study with participants randomly assigned to one of the groups in section 3.2. Fig. 2 shows a flowchart of the study design.

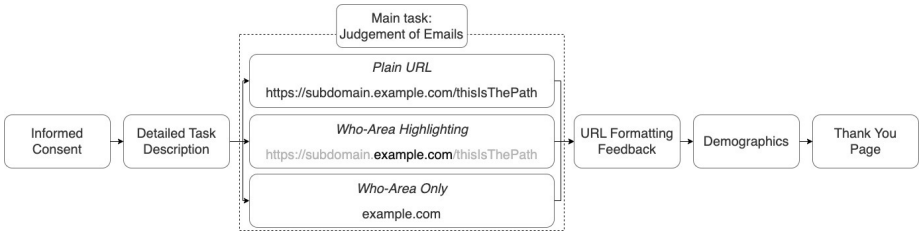

Fig. 2. Flowchart of the study design with examples of each group URL Formatting.

Informed Consent. We informed the participants of the study goal and their rights, including the data anonymous collection and analysis, and that they could stop the study at any time, without reasons, leading to their data deletion.

Detailed Task Description. We showed participants a detailed task description. We also informed them of the one second delay and we asked them to judge the emails as Martin Müller, who knows and uses all the services shown.

Main Task: Judgment of Emails. Participants saw twelve interactive email screenshots in the Gmail environment (see section 3.5), displayed in random order. Each screenshot was on a separate page with the question "On which web page does Martin land when he clicks on the link in the email? ... A phishing web page / A web page from *organization*".

URL Formatting Feedback. Participants were asked two separate open questions on what they liked about the formatting they saw and to mention one improvement they would like to see added to it. The specific formatting was shown in an image to avoid relying on the participants' memories.

Demographics. We asked participants their gender, age and if they received previous awareness material on phishing.

Thank You Page. We thanked participants for the participation and provided their code for payment (see section 3.6).

3.4 URL Techniques

An overview of the URLs used in the study is in Table 2. Volkamer et al. [36] and Xiong et al. [43] used the URL techniques by Lin. et al. [16]. For comparability, we also used the categorization from Lin et al. [16].

Obfuscate: An arbitrary domain name or IP address that hides the destination. The URL is not related or in any way similar to the impersonated sender of the email content. For example, in a phishing Amazon email, the URL behind a link can be `www.host745.com` or `https://87.147.12.250`.

Mislead: The impersonated sender company name is used as subdomain or added to the path. For example, in a phishing Amazon email, the URL behind a link can be `www.amazon.de.host745.com` or `www.host745.com/www.amazon.com`.

Mangle: The name of the impersonated sender company is used in the domain but with small, difficult to spot changes. For example, in a phishing Amazon email, the URL behind a link can be `www.amazno.com` (two characters inverted) or `www.arnazon.com` (using"`r`" "`n`" instead of "`m`").

We also excluded other techniques: mismatch attacks, short-URL and redirection services, and programmed tooltips.

A mismatch attack uses a link-text resembling a legitimate URL of the impersonated organization. Yet, the destination URL is different than the link-text, e.g., the link-text reads `www.amazon.de`, but the URL behind it is `www.evil.com`.

This technique can be combined with all three of Lin et al.'s techniques, introducing a different dimension than just URL manipulation. Hence, we acknowledge its existence, but postpone to future studies its investigation.

Short-URL services hide the actual destination URL, making it impossible to judge it without visiting the destination or using a tool to show the destination URL. Attacks using short-URL services should be solved by technical means and not rely on users' judgment, which cannot be based on informed decisions. Therefore, we decided to exclude them.

Redirection services work similarly to short-URL services, but the destination URL can usually be found in the path of the redirection URL. Yet, since many users are unaware of this, we decided to exclude them. Also, just like short-URL services, technical means should be employed to solve these cases.

Programmed tooltips use front-end manipulation (e.g., via JavaScript or CSS) to create fake tooltips showing whatever the attackers want. As in the previous two cases, these attacks should be addressed at the technical level.

Table 2. Overview of the phishing techniques matched to the companies and the URLs used in the study.

Company	Strategy Type	URL Used
Amazon	Obfuscate	`https://telefon.host745.com/hinzufuegen`
Lufthansa	Obfuscate	`https://87.147.12.250/buchungsänderung`
Google	Mislead	`https://www.google.com.megahoust.ru/sicherheitscheck`
LinkedIn	Mislead	`https://login.linkyzt.com/www.linkedin.com/profil`
DHL	Mangle	`https://account.dlh.com/zustellung`
Netflix	Mangle	`https://www.netfllx.com/neuerlogin`
Amazon	Legitimate	`https://packet.amazon.de/paketverfolgung`
Lufthansa	Legitimate	`https://www.lufthansa.com/buchungsanzeige`
Google	Legitimate	`https://www.google.com/neuesgerät`
LinkedIn	Legitimate	`https://video.linkedin.com/kurs`
DHL	Legitimate	`https://mailing.dhl.de/wunschort`
Netflix	Legitimate	`https://www.netflix.com/neuepreise`

3.5 Interactive Email Screenshots

We considered all phishing techniques in section 3.4 twice, leading to six phishing email screenshots and six legitimate email screenshots (twelve in total). This ratio was used before in similar studies (e.g., in [4, 6, 18, 27]). We chose to cover each technique twice to reduce the influence of both participants' biases (e.g., personal opinions on the organizations) and unexpected issues with the specific email screenshots used (e.g., confusing URL/wording). Each screenshot was interactive and showed a tooltip with the appropriate URL formatting once the link was hovered with the mouse cursor. The URL formatting depended on the study group (see section 3.2).

Furthermore, participants could not proceed to the next email screenshot if the tooltip was not continuously hovered for at least one second and we explicitly informed them of this in the Detailed Task Description step. Volkamer et al. [36] used a three second timer to block the clicking of links so to force users' attention on the URL formatting. However, the three seconds forced attention effect was not evaluated, i.e., a longer or a shorter period of time might have been better. Jain et al. [14] showed that humans require 250 ms on average to notice a visual cue. Yet, we want our participants to also read the tooltip content, not only to notice it appeared. Hence, we decided to extend the waiting time to one second to force our participants to read the tooltip content.

Regarding the emails screenshots, we used real world emails from well-known companies and we only replaced the URL behind the single link present. That is, we created advanced phishing emails as described in section 1.

The study was in German with German participants, hence, we chose companies highly popular in Germany (see Table 2), to reduce false results based on unfamiliarity with the sender.

3.6 Ethics, Recruitment and Payment

3.6.1 Ethics. The study design was reviewed and approved by the ethical board of our university. The informed consent was developed in collaboration with the data protection officer of our university and approved by the ethical board.

3.6.2 Power Analysis. We conducted a power analysis to determine the number of participants to recruit. As we wanted to compare the phishing detection of three groups, we planned to conduct a one-way ANOVA to evaluate the presence of difference, followed by a post-hoc analysis to determine if the difference was present in all pairs considered in our hypotheses. To control the family-wise error rate due to conducting multiple hypotheses tests (discussed in Miller [17]), we planned to employ a Tukey's test (detailed in Tukey [33]) during the post-hoc analysis.

We used G*Power[2] to conduct the power analysis. We set the test to ANOVA (fixed effects, omnibus, one-way), the α-error to .05 (as recommended in Fisher [13]) and the β-error to .90 (following Cohen [7]). Regarding effect size, Volkamer et al. [36] did not report their effect size, and Xiong et al. [43] did not conduct statistical analysis of intra-phase results (only inter-phases). Neither study reported standard deviation or mean of their results, making it impossible to even estimate the effect size they found. For this reason, we aimed at a medium effect size (Cohen's $r = .30$), as suggested by Cohen [8] when no justifications to expect a different effect size are available.

The parameters above could be met with 207 participants. Yet, to avoid issues due to drop-out or participants exclusion, we added a buffer of 5 participants per group. Hence, we recruited 222 participants.

3.6.3 Recruitment. We used the panel service "Clickworker" to recruit our participants, limiting the selection to those from Germany and speaking German. Only 210 participants completed the survey, instead of the expected 222. Twelve participants were not recruited due to technical problems outside of our control.

3.6.4 Payment. We wanted to pay our participants at the German minimum wage, i.e., € 12.00/h. Considering that our survey took 15 minutes to complete, each participant received € 3.00 ($12/60 * 15 = 3$).

4 RESULTS

In this section, we present the results with respect to the different research questions and hypotheses.

4.1 Data Cleaning

We performed the following data cleaning steps on the 210 participants' data:

(1) We excluded one participant as one of their hovering events was less than 1 second, concluding that it would be impossible for them to pay proper attention to the tooltip.
(2) We excluded nine participants as they consistently selected one option for every email, i.e., they marked either all emails as legitimate or all emails as phishing. Hence, we concluded that these participants did not pay attention to their task.

Thus, we considered the data of 200 participants. While this is less than the minimum 207 required to reach a power of .90 (see section 3.6.2), we still achieve the power of .89.

[2]Power analysis tool for tests, https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

4.2 Demographics

We report here an overview of the demographics of our participants. The demographics factors and the presence or absence of influence from them on the participants' URL detection rate (be it phishing, legitimate or overall) are explored in sections 4.5, 4.6 and 4.7. Table 3 provides the age distribution across the study groups. Table 4 shows the distribution of genders and previous phishing awareness across the study groups.

Table 3. Age-group distribution for all three groups.

Group	18-24	25-34	35-44	45-54	55-64	65+
Plain URL	10	20	22	12	6	4
Who-Area Highlighting	9	21	21	10	4	3
Who-Area Only	5	24	14	5	8	2

Table 4. Gender and previous phishing awareness distribution for all three groups.

Group	Male		Female		Don't say		Awareness		No Awareness	
Plain URL	48	64.86%	26	35.14%	-	-	40	54.05%	34	45.95%
Who-Area Highlighting	42	61.77%	25	36.76%	1	1.47%	37	54.41%	31	45.59%
Who-Area Only	40	68.97%	18	31.03%	-	-	29	50.00%	29	50.00%

4.3 Phishing URL, Legitimate URL and URL Overall Detection

The descriptive results in Table 5 seem to follow the pattern suggested by the results of Xion et al. [43] and Volkamer et al. [36], i.e., Plain URL is the worst performing group while Who-Area Only is the best. Table 6 shows that the worst performing URL technique is Mangle, which is considerably lower than the others.

Table 5. Percent of hit and correct rejection rate of the study groups divided between phishing and legitimate.

	Group	Mean	SD
Phishing	Plain URL	71.40%	0.25
	Who-Area Highlighting	73.53%	0.22
	Who-Area Only	76.44%	0.24
Legitimate	Plain URL	88.06%	0.17
	Who-Area Highlighting	91.42%	0.17
	Who-Area Only	93.10%	0.13

Table 6. Percent of hit and correct rejection rate of the study groups by URL technique.

URL Technique		Company	Phishing			Legitimate		
			Plain URL	*Who-Area Highlighting*	*Who-Area Only*	*Plain URL*	*Who-Area Highlighting*	*Who-Area Only*
Obfuscate	Arbitrary	Amazon	93%	94%	93%	84%	91%	97%
	IP Address	Lufthansa	78%	86%	76%	91%	90%	83%
Mislead	Subdomain	Google	85%	82%	93%	92%	93%	97%
	Path	LinkedIn	84%	84%	88%	89%	93%	90%
Mangle	Inversion	DHL	46%	43%	59%	84%	93%	98%
	Similarity	Netflix	42%	54%	50%	89%	90%	95%

To answer RQ1, we decided to evaluate if the difference in the phishing URL detection rate shown in Table 5 is statistically significant or not. First, we checked the parametric assumptions of the phishing URL detection data by conducting a Shapiro-Wilk test and a Levene's test. The Shapiro-Wilk test shows that the phishing URL detection data violated the normality assumption ($p < .001$), while the Levene's test shows that it meet the homogeneity of variance assumption ($p > .05$). Since the data only violates the normal distribution assumption, and because Field [12] points out that the ANOVA test is relatively robust against it, we continued the analysis using the ANOVA test. The results show no significant difference regarding the phishing URL detection ($F(2) = .73, p > .05$) with a $\eta^2 = 0.007$.

Hence, the data fails to reject the null hypotheses associated with $H1_{WAH>PL}$, $H2_{WAO>WAH}$, and $H3_{WAO>PL}$ from section 3.1. Thus, the answer to **RQ1** is that *no URL formatting is significantly better than the others at supporting users' phishing URL detection for our sample size.*

As we wanted to see if there was any difference in the legitimate URL detection data, we decided to analyze this data too. We checked again the parametric assumptions with a Shapiro-Wilk test and a Levene's test and they showed that the legitimate URL detection data violates both the assumptions of normality ($p < .001$) and homogeneity of variance ($p = .0109$). For this reason, we used a Kruskal-Wallis rank sum test to determine the presence of differences among the groups. The results show no significant difference regarding the legitimate URL detection ($\tilde{\chi}^2 = 2.65, p > .05, df = 2$).

Several studies on phishing (e.g., [9, 27, 34]) also consider the overall detection rate of participants. Hence, we decided to include this aspect too. The URL overall detection rate data violates the normality assumption (Shapiro-Wilk test, $p < .001$), but meet the homogeneity of variance assumption (Levene's test, $p > .05$). We ran an ANOVA test to check for differences, but found none ($F(2) = 1.66, p > .05$).

4.4 Hovering Time Influence

We divided our sample in quartiles according to the hovering time (shown in Table 11 in the appendix). The benefits on URL detection rate of longer hovering time seem to plateau between 3.97s and 5.72s; as for the fourth quartiles, there is no improvement for the overall number of correct answers.

To answer RQ2, we evaluated if there was a significant difference between the hovering time of the groups. We checked the parametric assumption with a Shapiro-Wilk test ($p < .001$) and a Levene's test ($p > .05$), finding that the data violated the normality assumptions, but not the homogeneity of variance one, allowing us to run an ANOVA test. We found a difference between groups ($F(2) = 3.59, p = .0293$) and run a Tukey Multiple Comparison of Means to determine which groups showed differences, founding that only the pair Who-Area Only and Plain URL did ($p = .0249$).

To determine if the influence of hovering time was different for phishing URL, legitimate URL and URL overall detection rate, we ran three separate ANCOVA test. The detailed output is shown in three tables in the appendix.

First, we ran an ANCOVA comparing the influence of hovering time on the study groups phishing URL detection, but we found no significance (Study groups, $F(2) = 2.39, p > .05$; Hovering time, $F(1) = 25.05, p < .001$). We then ran an ANCOVA comparing the influence of hovering time on the study groups legitimate URL detection, finding no significance either (Study groups, $F(2) = 2.11, p > .05$); Hovering time, $F(1) = 2.15, p > .05$). Finally, we ran an ANCOVA comparing the influence of hovering time on the study groups URL overall detection, finding a significant difference (Overall, $F(2) = 3.62, p = .0286$; Hovering time, $F(1) = 20.16, p < .001$). We then run a Tukey Contrasts test to determine which groups showed differences and found that only the pair Who-Area Only and Plain URL showed a difference ($p = .0211$).

Using a linear regression, we found a significant relationship between hovering time and URL overall detection rate ($R^2 = 0.075, F(1, 198) = 16.14, p < 0.001$).

We also checked if there was a correlation between the URL technique used and the hovering time based on the group. We again used linear regression and found significant relationships between hovering and Obfuscate detection ($R^2 = 0.083, F(1, 198) = 18.06, p < 0.001$), Mislead detection ($R^2 = 0.035, F(1, 198) = 7.16, p = 0.008$), and Mangle detection ($R^2 = 0.051, F(1, 198) = 10.77, p = 0.001$).

Regarding **RQ2**, then, we found three results: (1) *there is a significant difference in hovering time only between Who-Area Only (shorter) and Plain URL (longer)*; (2) *even if Who-Area Only has a significantly shorter hovering time than the Plain URL group, the URL overall detection rate of the first is better than the URL overall detection rate of the second*; (3) *there is a correlation between a longer hovering time and a better URL overall detection rate, and between a longer hovering time and the detection of each URL technique used.*

4.5 Gender Influence

As shown in Table 4, only one participant selected the "Don't say" option regarding gender. Because a single person group would confuse the analysis, we decided to exclude them from the gender influence analysis. In other words, only for the analyses reported in this section, our sample is 199 participants, not 200. The participant is considered in any other analysis besides the gender influence one.

We checked the parametric assumptions of the remaining gender groups and they violated both the normality assumption (Shapiro-Wilk test, men, $p < .001$; women, $p = .002$), and the homogeneity of variance assumption (Levene's test, $p = .0128$). We ran a Welch Two Sample t-test and found significant difference ($t(113.62) = 3.78, p < .001$) between the URL overall detection rate of the gender groups, with a medium effect size $d = .58$.

In addition to the basic difference between the gender, we now wanted to look at the difference between the two URL formatting regarding gender as a further investigation (see Table 7). To do this, we first looked at the descriptive values for the two genders divided among the three groups. Here, it iss noticeable that both men and women showed an improvement in the number of correct answers for both URL formatting. Yet, men benefit more from Who-Area Only (averaging just over 5.1%) than Who-Area Highlighting (averaging just over 1.8%). In contrast, women benefit from both representations - but, unlike men, they benefit more from Who-Area Highlighting (averaging just over 4.5%), than from Who-Area Only (averaging 3.4%).

Table 7. Difference of percentage of correct answers for group and gender factors.

Group	Gender	N	Mean % of correct answers	Improvement from Plain URL to	SD	Min	Max
Plain URL	men	48	83.17	-	13.83	41.67	100
	women	26	73.42	-	19.42	33.33	100
Who-Area Highlighting	men	42	85.00	1.83	15.75	41.67	100
	women	25	78.00	4.58	14.58	41.67	100
Who-Area Only	men	40	88.33	5.17	11.75	41.67	100
	women	18	76.83	3.42	19.50	25	100

4.6 Age Influence

We checked the normality assumption with a Shapiro-Wilk test ($p < .001$), but it was violated. We then checked the homogeneity of variance assumption with a Levene's test ($p > .05$), which was met. We ran an ANOVA test to see if there was a significant difference in the URL overall detection rate by the age groups, but we found none ($F(5) = .28, p > .05$).

4.7 Previous Phishing Awareness Influence

We ran a Shapiro-Wilk test to check the normality assumption and it was violated (Aware, $p < .001$; Not Aware, $p < .001$). The data met the homogeneity of variance (Levene's test, $p > .05$). We ran a Two Sample t-test but found no significant difference in the URL overall detection rate by previous phishing awareness ($t(198) = -1.51, p > .05$).

4.8 Qualitative Data

As mentioned in section 3.3, we asked our participants two open questions: what did they like of the URL formatting they saw during their main task and to mention one aspect we should improve of it. Yet, we did not find the qualitative results neither interesting nor illuminating towards our research questions. For this reason, we decided to not consider them in this paper and we plan to investigate them more thoroughly in a future study.

5 DISCUSSION

This section discuss the implications of our results.

5.1 URL Formatting Effectiveness

Determining the identity of a specific web page based on the URL is hard, and this is a renown fact: the Google Chrome team decided to phase out the use of URLs due to security concerns, as reported in Newman [21]. This was further mentioned by Emily Stark, head of Chrome "Trusty Transport" team, in a comment to a bug report thread of "The Chromium Project" (see Stark [30]). Yet, finding a way to support users is also quite hard, as shown roughly one year later by Stark herself, who mentioned that their "simplified domain experiment" was not effective, according to their security metrics, *de facto* stopping it (see Stark [31]).

Nonetheless, it seems that at least some industry actors still have trust in the benefits of URL formatting: Apple Safari on MacOS, in its "Smart Search Field", uses by default a URL formatting similar to Who-Area Only. Only explicitly ticking an option in the "Advanced Preferences" pane allow the users to see the full URL, as explained on the official Apple support web page (see Apple [32]). Mozilla Firefox developers too decided to apply URL formatting by default, namely, Who-Area Highlighting. The option to use it or not (a boolean function named "browser.urlbar.formatting.enabled") is effectively hidden from the UI normally accessible to users, as mentioned in the Firefox documentation (see Mozilla [20]).

We found the same apparent pattern with regard to phishing URL detection as mentioned in section 2, i.e., Who-Area Only have the best performance (76.44%) followed by Who-Area Highlighting (73.53%) and Plain URL is the worst one (71.40%). Yet, the inferential statistics showed no statistical significance to this pattern within our sample size (see section 4.3). Hence, with respect to RQ1, for our sample size URL formatting not only had no significant effect on phishing URL detection, but it also had no significant effect on legitimate URL detection and URL overall detection.

Admittedly, it can be that our sample size was not big enough to find a significant effect. Yet, we want to highlight three points in this regard: the first one is that we aimed at a medium effect, conducting a power analysis and recruiting enough participants to detect one. As shown in section 4.3, the only effects we detected were small ones, so a significant effect would most likely be such. The second point is that we applied the URL formatting in the participants' attention focus using tooltips. That is, according to several related work results (e.g., [25, 36, 43]), if URL formatting had a non-small significant effect on phishing URL detection, it should have been more easily detected in our setting than in the real-world. This because, in the latter, URL formatting is not normally in the users' attention focus (e.g., it is shown in a web address bar). Thirdly, in our setting, security was the participants' main task. In everyday life, users do not

check emails for phishing URLs in this way, it is a secondary task that happens while carrying out the main one. This mean that, in our setting, participants were more focused and more influenced by factors that would influence phishing URL detection. Considering all three of our points, especially the last two, makes our setting a "best-case" scenario to detect a significant non-small URL formatting effect. Yet, even in this best case scenario no such effect was found.

Nonetheless, even if the security value of URL formatting is debatable, it might still have usability benefits, due to the reduced number of URL parts shown to users. Yet, this potential usability benefit assumes that users are indeed capable of reading and understanding URLs, which several studies (e.g., [1, 24, 45]) have shown not to be the case. However, the usability aspect of URL formatting requires further and dedicated investigation.

5.2 Hovering Time Influence

Our results show that the only difference present in hovering time is between Who-Area Only and Plain URL. Considering the difference in the length of what is shown between Plain URL (longer) and Who-Area Only (shorter), this is not so surprising. What is surprising is the lack of difference between Who-Area Highlighting and Plain URL: the presence of highlighting should reduce the time required to check a URL by directing the users' to the highlighted part. Yet, our results suggest that simply highlighting the who-area is not sufficient to significantly decrease the hovering time.

Regarding the influence of hovering time on the participants' URL detection rate, the correlation between the two suggests that users should take their time when checking URLs, as it demonstrably leads to a better phishing URL detection. It is particularly interesting that, even though Who-Area Only has a significantly shorter hovering time than Plain URL, the first have a significantly better URL detection rate than the latter. Hence, even though URL formatting alone is not influential on the users' phishing URL detection rate, it might still be beneficial to adopt a shorter URL formatting, as it requires the lowest time investment while leading to a better URL overall detection rate.

It could be said that reaching a shorter time investment is especially important in the business context, where longer times correspond to a loss in productivity. However, this is a simplification that we do not fully support. Albeit it is true that "time is money" in the business context, it is important to point out that security is usually not the main goal of users irrespective of the context. Hence, given the potential for security fatigue, i.e., ignoring some time-intensive security recommendations to reach one's goal (as described in Stanton et al. [29]), reducing the time required to successfully complete a security task is beneficial in any context.

It should be noted, though, that a potential influencing factor in the hover time dynamic is the use of tooltips. As mentioned throughout the paper, we specifically selected tooltips due to the attention grabbing effect described in Petelka et al. [25] and the beneficial effects on URL detection rate that attention has. If the URL formatting is not in the center of attention, it might be that the beneficial effect of time on phishing URL detection would not be present due to users not noticing the URL formatting at all (because outside of their attention focus). However, more study is required to confirm this dynamic and we plan as future work to further investigate the time effect on phishing URL detection.

5.3 Further Influential Factors

An interesting result is that the simple consideration of the URL overall detection results, including groups and hover time, led to a clear result (Who-Area Only would be the best methodology). However, when the results are now broken down to gender, the picture is no longer so clear. The question now arises as to which factors have an influence (like gender) or if the difference found actually exists. This raises the question of whether there should be a clear tendency towards a single URL formatting or a differentiated one based on other factors. For example, the differences found may not be based on gender, but on correlated personality traits such as conscientiousness or cognition (Delaney et

al. [10]'s rational vs. intuitive thinking or Putrevu [26]'s selectivity interpretation). Because the focus of our study was not primarily on answering these questions, follow-up studies should place a stronger focus here. We especially want to point out that we have not controlled the demographics of our sample strongly enough to make any definitive statement on the influence of gender on URL formatting. As our goal was not to evaluate differences due to gender, personality, cognition, and similar, the study was not designed to carefully check every potential influence. Our sample is also unequally distributed among genders. Although the statistical tests we used are robust against such inequality, it is important to follow-up our work with a dedicated evaluation better suited to investigate these influences. Our results, however, hint that there might be some interaction between gender (or related factors) and URL formatting. As future study, we plan to design and run a dedicated study to further investigate these aspects.

Also interesting is that neither age nor previous phishing awareness had any significant influence on the participants' URL detection rate. The lack of age influence might be due to the unbalance between age groups shown in Table 3. It is then necessary to explore the age influence in dedicated future work.

Regarding previous phishing awareness, though, the distribution among the groups was fairly even (see Table 4). The reason for the lack of influence may be found in the results of Wash et al. [40], who reports that only 44% of their participants notice links in emails, meaning that 56% do not check the URL. Yet, Zheng & Becker [45] found that even those users that do check the URL behind a link, lack knowledge regarding what to do with this information. The results confirmed those of previous studies, e.g., Albakry et al. [1] and Pearson et al. [24], who showed how users lack understanding of how to read URLs. Furthermore, the results of Mossano et al. [19] showed that publicly available anti-phishing web pages are not following a standardized set of recommendations, potentially causing an uneven level of awareness. As, presumably, our participants did not receive their awareness from the same source, they might have different information, especially regarding URL analysis. Further studies on the relation between anti-phishing awareness and URL formatting are required.

5.4 Impact on the Wider Security Landscape

So far we have discussed our results separately. We would now like to condense our insights in a cohesive narrative to elaborate on their impacts on the wider security landscape, especially on how the security community could act on them. As discussed in section 5.1, URL formatting is currently used in several programs, chief of all in browsers. This is because it is assumed that URL formatting has a beneficial effect on phishing URL detection. Our results show that this is not the case, not even in a "best case scenario" where the users' focus is on security and URLs. The only significant effect we found appears when users spend time on URLs, but even in that case, the effect is suggested to be different depending on gender-related factors. Given all of the above, our message to the wider security community is four folds:

- As our results show that hovering time has a beneficial effect on phishing URL detection, we suggest to focus on interventions that incorporate this, e.g., the TORPEDO add-on from Volkamer et al. [38].
- As our results hint at the presence of a ceiling to the beneficial effect of hovering time, we suggest to systematically investigate this aspect in a dedicated study and to integrate the results in the interventions developed in the future.
- As our and other studies (e.g., [1, 24, 45]) suggest that users lack knowledge on what to do with the URL information, we suggest to expand the current awareness measures or to develop new ones to address this aspect.
- As our results hint at the presence of gender-related influential factors on URL formatting, we suggest that these are studied more systematically in a dedicated study.

5.5 Limitations

Like every other study, ours too have limitations in its methodology.

First of all, phishing detection is not usually the main task for users. Having security as main task creates an unrealistic focus on it, which would not be the case in the real world. Yet, although the external validity of our results might be debatable given the unrealistic setting, it is important to note that our results show no significant non-small effect in what is, effectively, a best case scenario. By this we mean that, if our results show no significant non-small effect from URL formatting in a setting that is more likely to have it than the real world, it is doubtful that the effect would be significant in a setting were the users' attention is focused on other tasks. Hence, we believe that the security focus does not decrease our results importance.

The second limitation of our study is that we only considered a small part of phishing detection, i.e., the URL analysis. Nonetheless, as our focus was on URL analysis, we do not think this decreases our results importance.

Furthermore, we used interactive screenshots instead of showing the emails in a functioning email environment. This might have introduced changes in the participants' behavior, as their interactions were limited to those allowed by the screenshots themselves: for example, they could not hover the sender address, but only the link in the text. Yet, the sender address was shown in full and it was always the legitimate address of the legitimate organization, i.e., no cue about the email legitimacy could be inferred from it.

We also introduced a one second forced hovering of the tooltip. We did so to make sure participants took their time to read the URL shown in the tooltip and to answer one of our research questions. Although this might have modified the participants' normal behavior, given Petelka et al. [25]'s results that forced attention increases phishing URL detection, our choice should have increased the participants' phishing URL detection and the effect of URL formatting. As this did not happened, we do not believe that the one second forced hovering time had any harmful effect on our results.

We considered only those URL formatting found in the literature. Although this was done for comparative reasons, it might be the case that other URL formatting not considered would lead to different results.

We did not control further influential factors related to demographics, such as current occupation, previous education and daily interaction with emails. We also did not control the gender balance of our sample. This might have introduced imbalances in our gender-related results, which should be properly addressed in a follow-up study.

Our study was localized in Germany, with German participants. A different population might lead to different results.

6 CONCLUSION

Our goal was to determine if URL formatting had any influence on the phishing URL detection rate of users. We designed a between-subjects, online study with 200 German participants and evaluated two URL formatting from related work (Who-Area Highlighting and Who-Area Only) against a control group with unmodified URLs (Plain URL). Our results showed no statistical difference among the URL formatting. Yet, we found that other factors might have an influence on the participants' URL detection: the time spent hovering the link to show the tooltip and gender-related factors. Albeit the security benefit of URL formatting on its own seems limited, further studies are required to determine if, when used in combination with other factors, there might be a case to propose its use, e.g., for usability.

ACKNOWLEDGMENTS

This work was supported by funding from the project "Engineering Secure Systems" of the Helmholtz Association (HGF) [topic 46.23.01 Methods for Engineering Secure Systems] and by KASTEL Security Research Lab.

REFERENCES

[1] Sara Albakry, Maria K. Wolters, and Kami Vaniea. 2020. What is this URL's Destination? Empirical Evaluation of Users' URL Reading. In *Conference on Human Factors in Computing Systems* (Honolulu, HI, US) *(CHI '20)*. ACM, New York, NY, US, 1–12. https://doi.org/10.1145/3313831.3376168

[2] Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. 2015. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies* 82 (2015), 69–82. https://doi.org/10.1016/j.ijhcs.2015.05.005

[3] APWG. 2022. *Phishing Activity Trends Report.* Technical Report. Anti-Phishing Working Group. Issue: 3rd quarter 2022.

[4] Benjamin M. Berens, Katerina Dimitrova, Mattia Mossano, and Melanie Volkamer. 2022. Phishing awareness and education – When to best remind?. In *Usable Security and Privacy Symposium* (San Diego, CA, US) *(USEC '22)*. Internet Society, Reston, VA, US. https://doi.org/10.14722/usec.2022.23075

[5] Vaishnavi Bhavsar, Aditya Kadlak, and Shabnam Sharma. 2018. Study on Phishing Attacks. *International Journal of Computer Applications* 182, 33 (2018), 27–29. https://doi.org/10.5120/ijca2018918286

[6] Casey Canfield, Baruch Fischhoff, and Alex Davis. 2015. Using Signal Detection Theory to Measure Phishing Detection Ability and Behavior. In *11th Symposium on Usable Privacy and Security* (Ottawa, ON, CA) *(SOUPS '15)*. USENIX, Berkeley, CA, US. https://cups.cs.cmu.edu/soups/2015/posters/soups2015_posters-final13.pdf

[7] Jacob Cohen. 1992. A power primer. *Psychological Bulletin* 112, 1 (1992), 155–159.

[8] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences.* Academic Press, New York, NY, US.

[9] Sanchari Das, Christena Nippert-Eng, and L. Jean Camp. 2022. Evaluating User Susceptibility to Phishing Attacks. *Information and Computer Security* 30, 1 (2022), 1–18. https://doi.org/10.1108/ICS-12-2020-0204

[10] Rebecca Delaney, JoNell Strough, Andrew M. Parker, and Wandi Bruine de Bruin. 2015. Variations in decision-making profiles by age and gender: A cluster-analytic approach. *Personality and Individual Differences* 85 (2015), 19–24. https://doi.org/10.1016/j.paid.2015.04.034

[11] Rachna Dhamija, J. Doug Tygar, and Marti Hearst. 2006. Why Phishing Works. In *Conference on Human Factors in Computing Systems* (Montréal, QC, CA) *(CHI '06)*. ACM, New York, NY, USA, 581–590. https://doi.org/10.1145/1124772.1124861

[12] Andy Field. 2013. *Discovering Statistics Using IBM SPSS Statistics.* Sage Publications, Thousand Oaks, CA, US.

[13] Ronald A. Fisher. 1921. On the Probable Error of a Coefficient of Correlation Deduced from a Small Sample. *Metron* 1 (1921), 3–32.

[14] Aditya Jain, Ramta1 Bansal, Avnish Kumar, and KD Singh. 2015. A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. *International Journal of Applied and Basic Medical Research* 5, 2 (2015), 124–127. https://doi.org/10.4103/2229-516X.157168

[15] Kaspersky. 2023. *Spam and phishing in 2022.* Kaspersky. Retrieved June 9, 2023 from https://securelist.com/spam-phishing-scam-report-2022/108692/

[16] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. 2011. Does domain highlighting help people identify phishing sites?. In *Conference on Human Factors in Computing Systems* (Vancouver, BC, CA) *(CHI '11)*. ACM, New York, NY, US, 2075. https://doi.org/10.1145/1978942.1979244

[17] Rupert G. Miller. 1981. *Simultaneous Statistical Inference.* Springer, New York, NY, US.

[18] Mattia Mossano, Benjamin M. Berens, Philip Heller, Christopher Beckmann, Lukas Aldag, Peter Mayer, and Melanie Volkamer. 2022. SMILE - Smart eMaIl Link Domain Extractor. In *ESORICS 2021 International Workshops* (Darmstadt, DE) *(ESORICS '21)*. Springer, Cham, 403–412. https://doi.org/10.1007/978-3-030-95484-0_23

[19] Mattia Mossano, Kami Vaniea, Lukas Aldag, Reyhan Düzgün, Peter Mayer, and Melanie Volkamer. 2020. Analysis of publicly available anti-phishing webpages: contradicting information, lack of concrete advice and very narrow attack vector. In *European Symposium on Security and Privacy Workshops* (Genoa, IT) *(EuroS&PW '20)*. IEEE, New York, NY, US, 130–139. https://doi.org/10.1109/EuroSPW51379.2020.00026

[20] Mozilla. 2023. *Preferences.* Mozilla. Retrieved June 9, 2023 from https://firefox-source-docs.mozilla.org/browser/urlbar/preferences.html

[21] Lily H. Newman. 2018. *Google Takes Its First Steps Toward Killing the URL.* Wired. Retrieved June 9, 2023 from https://www.wired.com/story/google-chrome-kill-url-first-steps/

[22] James Nicholson, Lynne Coventey, and Pam Briggs. 2017. Can we fight social engineering attacks by social means? Assessing social salience as a means to improve phish detection. In *13th Symposium on Usable Privacy and Security* (Santa Clara, CA, US) *(SOUPS '17)*. USENIX, Berkeley, CA, US, 285–298. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/nicholson

[23] Kunal Pandove, Amandeep Jindal, and Rajinder Kumar. 2010. Email Spoofing. *International Journal of Computer Applications* 5, 1 (2010), 27–30. https://doi.org/10.5120/881-1252

[24] Ed Pearson, Cindy L. Bethel, Andrew F. Jarosz, and Mitchell E. Berman. 2017. "To click or not to click is the question": Fraudulent URL identification accuracy in a community sample. In *International Conference on Systems, Man, and Cybernetics* (Banff, AB, CA) *(SMC '17)*. IEEE, New York, NY, US, 659–664. https://doi.org/10.1109/SMC.2017.8122682

[25] Justin Petelka, Yixin Zou, and Florian Schaub. 2019. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings. In *Conference on Human Factors in Computing Systems* (Glasgow, SC, UK) *(CHI '19)*. ACM, New York, NY, US, 1–15. https://doi.org/10.1145/3290605.3300748

[26] Sanjay Putrevu. 2001. Exploring the Origins and Information Processing Differences Between Men and Women: Implications of Advertisers. *Academy of Marketing Science Review* 10 (2001). http://www.amsreview.org/articles/putrevu10-2001.pdf

[27] Benjamin M. Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Düzgün, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. 2020. An investigation of phishing awareness and education over time: When and how to best remind users. In *16th Symposium on Usable Privacy and Security* (Online) *(SOUPS '20)*. USENIX, Berkeley, CA, US, 259–284. https://www.usenix.org/conference/soups2020/presentation/

reinheimer

[28] Joshua Reynolds, Deepak Kumar, Zane Ma, Rohan Subramanian, Meishan Wu, Martin Shelton, Joshua Mason, Emily Stark, and Michael Bailey. 2020. Measuring Identity Confusion with Uniform Resource Locators. In *Conference on Human Factors in Computing Systems* (Honolulu, HI, US) *(CHI '20)*. ACM, New York, NY, US, 1–12. https://doi.org/10.1145/3313831.3376298

[29] Brian Stanton, Mary F. Theofanos, Sandra Spickard Prettyman, and Susanne Furman. 2016. Security Fatigue. *IT Professional* 18, 5 (2016), 26–32. https://doi.org/10.1109/mitp.2016.84

[30] Emily Stark. 2020. *Issue 1090393: Implement simplified domain display in the omnibox, Comment 16.* The Chromium Project. Retrieved June 9, 2023 from https://bugs.chromium.org/p/chromium/issues/detail?id=1090393#c16

[31] Emily Stark. 2021. *Issue 1090393: Implement simplified domain display in the omnibox, Comment 75.* The Chromium Project. Retrieved June 9, 2023 from https://bugs.chromium.org/p/chromium/issues/detail?id=1090393#c75

[32] Apple Support. 2023. *Change Advance settings in Safari on Mac.* Apple. Retrieved June 9, 2023 from https://support.apple.com/guide/safari/advanced-ibrw1075/16.1/mac/13.0

[33] John W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2 (1949), 99–114. https://doi.org/10.2307/3001913

[34] Ploy Unchit, Sanchari Das, Andrew Kim, and L. Jean Camp. 2020. Quantifying Susceptibility to Spear Phishing in a High School Environment Using Signal Detection Theory. In *Human Aspects of Information Security and Assurance* (Mytilene, GR) *(HAISA '20)*. Springer, Cham, 109–120. https://doi.org/10.1007/978-3-030-57404-8_9

[35] Verizon. 2022. *Data Breach Investigations Report.* Technical Report. Verizon.

[36] Melanie Volkamer, Paul Gerber, and Karen Renaud. 2016. Spot the phish by checking the pruned URL. *Information and Computer Security* 24, 4 (2016), 372–385. https://doi.org/10.1108/ics-07-2015-0032

[37] Melanie Volkamer, Karen Renaud, Gamze Canova, Benjamin M. Reinheimer, and Kristoff Braun. 2015. Design and Field Evaluation of PassSec: Raising and Sustaining Web Surfer Risk Awareness. In *8th International Conference on Trust and Trustworthy Computing* (Heraklion, GR) *(TRUST '15)*. Springer, Cham, 104–122. https://doi.org/10.1007/978-3-319-22846-4_7

[38] Melanie Volkamer, Karen Renaud, and Benjamin M. Reinheimer. 2016. TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. In *31st IFIP TC 11 International Conference* (Ghent, BE) *(SEC '16)*. Springer, Cham, 161–175. https://doi.org/10.1007/978-3-319-33630-5_12

[39] Rick Wash. 2020. How Experts Detect Phishing Scam Emails. *Human-Computer Interaction* 4, CSCW2 (2020), 1–28. https://doi.org/10.1145/3415231

[40] Rick Wash, Norbert Nthala, and Emilee Rader. 2021. Knowledge and Capabilities that Non-Expert UsersBring to Phishing Detection. In *17th Symposium on Usable Privacy and Security (SOUPS '21)*. USENIX, Berkeley, CA, US, 377–395. https://www.usenix.org/conference/soups2021/presentation/wash

[41] Min Wu, Robert C. Miller, and Simson L. Garfinkel. 2006. Do security toolbars actually prevent phishing attacks?. In *Conference on Human Factors in Computing Systems* (Montréal, QC, CA) *(CHI '06)*. ACM, New York, NY, USA, 601–610. https://doi.org/10.1145/1124772.1124863

[42] Min Wu, Robert C. Miller, and Greg Little. 2006. Web wallet: preventing phishing attacks by revealing user intentions. In *2nd Symposium on Usable Privacy and Security* (Pittsburgh, PA, US) *(SOUPS '06)*. ACM, New York, NY, USA, 102–113. https://doi.org/10.1145/1143120.1143133

[43] Aiping Xiong, Robert W. Proctor, Weining Yang, and Ninghui Li. 2017. Is Domain Highlighting Actually Helpful in Identifying Phishing Web Pages? *Human Factors* 59, 4 (2017), 640–660. https://doi.org/10.1177/0018720816684064

[44] Weining Yang, Aiping Xiong, Jing Chen, Robert W. Proctor, and Ninghui Li. 2017. Use of Phishing Training to Improve Security Warning Compliance: Evidence from a Field Experiment. In *Hot Topics in Science of Security: Symposium and Bootcamp* (Hanover, MD, US) *(HoTSoS '17)*. ACM, New York, NY, USA, 52–61. https://doi.org/10.1145/3055305.3055310

[45] Sarah Zheng and Ingolf Becker. 2022. Presenting Suspicious Details in User-Facing E-mail Headers Does Not Improve Phishing Detection. In *18th Symposium on Usable Privacy and Security* (Boston, MA, US) *(SOUPS '22)*. USENIX, Berkeley, CA, US. https://www.usenix.org/conference/soups2022/presentation/zheng

APPENDIX

Table 8. Results of the ANCOVA test on the influence of the overall hovering time on the phishing detection

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	317.68	1	174.86	2.2^{-16}
Phishing Detection	8.70	2	2.39	0.0940
Time Overall	45.50	1	25.05	1.242^{-6}
Residuals	356.08	196		

Table 9. Results of the ANCOVA test on the influence of the overall hovering time on the legitimate detection

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	748.34	1	788.36	2.2^{-16}
Legitimate Detection	4.01	2	2.11	0.1237
Time Overall	2.04	1	2.15	0.1439
Residuals	186.05	196		

Table 10. Results of the ANCOVA test on the influence of the overall hovering time on the overall detection

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	2041.18	1	615.60	2.2^{-16}
Overall Detection	23.99	2	3.62	0.0287
Time Overall	66.83	1	20.16	1.217^{-5}
Residuals	649.89	196		

Table 11. Overall hovering time for the 12 examples divided into quartiles. The first column describes the 4 quartiles. The second column is the number of participants. The third column describes the overall lower bound of the quartile over all 12 examples and on average per example. The fourth column describes the higher bound of the quartile over all 12 examples and on average per example. The fifth and sixth columns describe again the mean number of overall correct answers (phish + legitimate) and its standard deviant.

Quartile	**Part.**	**Min. Time** (Avrg. Time)	**Max. Time** (Avrg. Time)	**Mean** ($max = 12$)	**SD**
0- 25%	50	16.04s (1.33s)	35.93s (2.99s)	8.78	2.37
25- 50%	50	35.93s (2.99s)	47.74s (3.97s)	9.6	1.59
50- 75%	50	47.74s (3.97s)	68.64s (5.72s)	10.5	1.67
75-100%	50	68.64s (5.72s)	194.13s (16.18s)	10.5	1.34

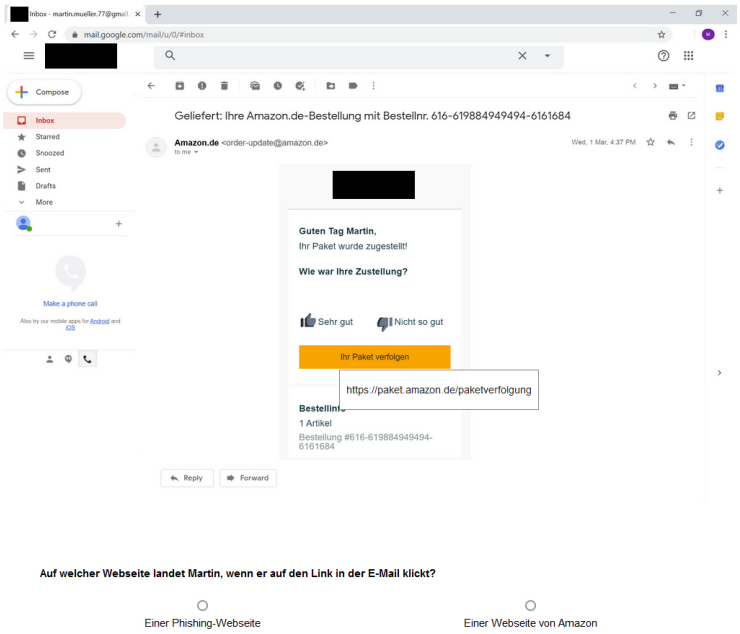

Fig. 3. Study interactive email showing the tooltip as seen by the participants in the Plain URL group

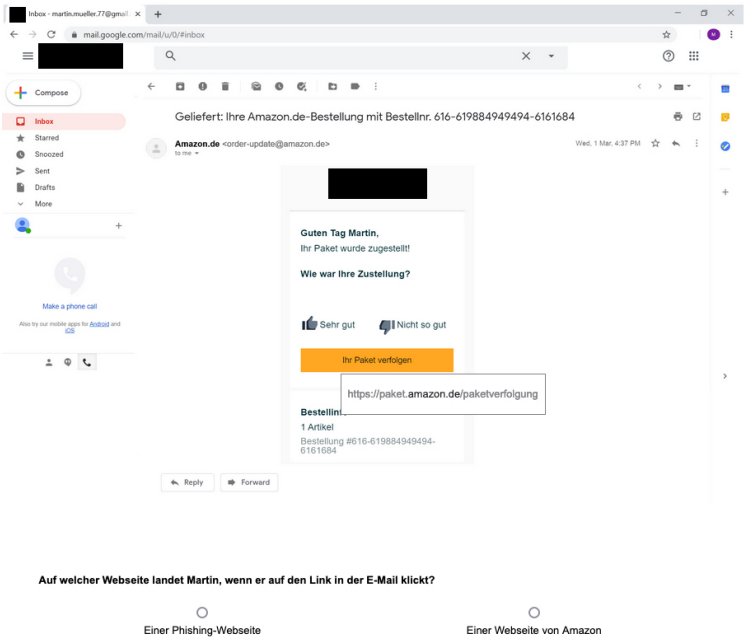

Fig. 4. Study interactive email showing the tooltip as seen by the participants in the Who-Area Highlighting group

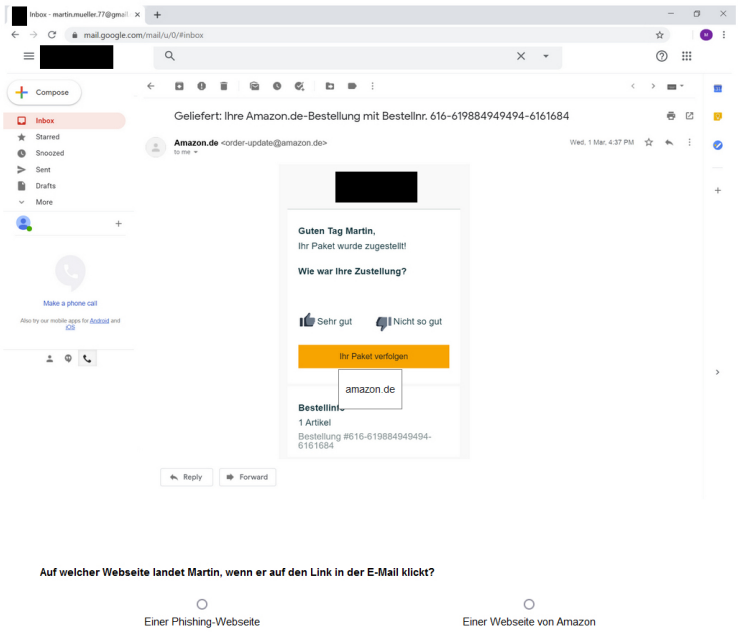

Fig. 5. Study interactive email showing the tooltip as seen by the participants in the Who-Area Only group