Vision: Retiring Scenarios — Enabling Ecologically Valid Measurement in Phishing Detection Research with PhishyMailbox

Oliver D. Reithmaier
Leibniz University Hannover
oliver.reithmaier@itsec.uni-hannover.de

Thorsten Thiel
Atmina Solutions
thorsten.thiel@atmina.de

Anne Vonderheide and Markus Dürmuth
Leibniz University Hannover
anne.vonderheide@stud.uni-hannover.de
markus.duermuth@itsec.uni-hannover.de

Abstract—**Email phishing to date still is the most common attack on IT systems. While early research has focused on collective and large-scale phishing campaign studies to enquire why people fall for phishing, such studies are limited in their inference regarding individual or contextual influence of user phishing detection. Researchers tried to address this limitation using scenario-based or role-play experiments to uncover individual factors influencing user phishing detection. Studies using these methods unfortunately are also limited in their ability to generate inference due to their lack of ecological validity and experimental setups. We tackle this problem by introducing PhishyMailbox, a free and open-source research software designed to deploy mail sorting tasks in a simulated email environment. By detailing the features of our app for researchers and discussing its security and ethical implications, we demonstrate the advantages it provides over previously used paradigms for scenario-based research, especially regarding ecological validity as well as generalizability through larger possible sample sizes. We report excellent usability statistics from a preliminary sample of usable security scientists and discuss ethical implications of the app. Finally, we discuss future implementation opportunities of PhishyMailbox in research designs leveraging signal detection theory, item response theory and eye tracking applications.**

I. Introduction

Phishing is one of the most common threats to IT systems worldwide. Malicious actors from ransomware gangs to state sponsored hacking groups use phishing emails as a first step to infiltrate target networks. Despite great efforts of researchers and practitioners, phishing in 2024 still is the most common cyberattack [1]. One of the reasons for this is, unfortunately, that popular measures like phishing simulation campaigns fail to deliver results [2]–[4]. In addition, phishing awareness campaigns trying to leverage education for higher security awareness also fall flat in terms of their effect — for an overview, see Sasse et al. [5]. While a relatively large body of literature regarding different influence factors on phishing detection exists, phishing attacks continue to evolve [6], [7], requiring better understanding of how and why email recipients detect phishing — or fall for it. Hence, research about phishing remains highly relevant to study and evaluate user behaviour as well as explore possibilities to strengthen users against such attacks.

Research regarding influences on phishing detection by users mainly employs one of two designs: While there is work implementing an observational perspective [8], most studies about influences on phishing detection have been operationalized using a scenario-based design, in which users view email images or isolated emails in combination with other measures like surveys, trying to elicit causal inference of said measures onto phishing detection performance [9]–[13].

Such scenario-based designs, however, show considerable drawbacks regarding content validity, ecological validity, inferential potential and overall generalizability. This warrants attention, as quantitative phishing detection research rests on its measurement. Since measures for behaviour must be ecologically valid to enable inference about the phenomenon it proposes to measure, measuring phishing detection so far has proven to be tough for researchers. To follow experimental best practices and maximize validity, researchers ought to use real phishing emails together with personal email clients. However, this approach conflicts with ethical guidelines prioritizing system health and privacy. Scenario-based designs by definition are an abstract solution to this, but they seldom if ever represent reality well. Either, participants are presented images of emails [9] or are shown emails in a locked-down inbox without being able to easily record user interaction with the email [14].

All of these approaches prevent actionable inference about phishing detection, no matter how well the study is run. The user either lacks familiar interaction, or there is no possibility to economically measure interaction. In this position paper, we detail three steps to improve research on human phishing detection and related constructs: Firstly, we carefully discuss limitations of scenario-based designs for phishing detection, based on extensive literature on the topic. We then introduce *PhishyMailbox*, a novel simulation software specifically developed to enable and measure ecologically valid user engagement with emails. Finally, we report preliminary usability data of researchers using PhishyMailbox and outline several use-cases for the software in future phishing detection research.

Symposium on Usable Security and Privacy (USEC) 2025
24 February 2025, San Diego, CA, USA
ISBN 979-8-9919276-5-9
https://dx.doi.org/10.14722/usec.2025.23037
www.ndss-symposium.org, https://www.usablesecurity.net/USEC/

II. Limitations of Quantitative Phishing Research

In this section, we provide an overview of quantitative phishing detection research, as well as methodological pitfalls of various research methods in that domain.

A. Limitations of Phishing Campaign Studies

Shortly after musings about phishing started appearing more prominently in academic circles in the early 2000s [15], Ferguson's work [16] — regarded as the first to use a large phishing campaign approach — found high rates of compliance with phishing emails in a military context; They effectively highlighted the importance of protection from phishing through training. Large-scale phishing campaign research, subsequently gained popularity despite known ethical concerns [17]: Dodge et al. [18] found similar phishing success rates to Ferguson [16] in a different military study. Jagatic et al. [19] first focused on social cues and concluded they contributed majorly to phishing success rates. Impact of demographics onto phishing detection was ruled out [20], and training of various types showed a positive effect on phishing detection [16], [21]–[23]. However, the impact of individual factors like psychological variables or exogenous influences on the users is impossible to obtain from such studies, as they are carried out in uncontrolled environments. While such large studies are ecologically valid, they cannot assess individual data well, as individual report metrics cannot be linked to study performance data.

To address this issue, researchers in the past decade increasingly adopted methods that allowed enquiring about influences on individual users through experimental manipulation in environments that allow for better measurement, i.e lab studies.

B. Scenario-based Designs

The scenario-based design rose in interest once the large-scale phishing studies became increasingly harder to conduct due to the spread of ethical concerns in the research community [17] and researchers considered individual factors as influence on phishing detection. A scenario-based design, often also called role-play design, is a study method first used in phishing detection by Downs et al. [24]. It describes immersing participants in scenarios in which they have to decide between alternatives. In phishing research, often a participant is told they are viewing their own email inbox or a similar email spectator event, to then be given the task of classifying emails into phishing or not, or evaluating them in terms of trust (cf. [9], [14], [24]).

The methodologies used in the literature can be broadly divided into two approaches: Either, researchers use email pictures to show participants — e.g. [25] — or prepare an isolated email inbox with pre-specified emails for the participants to handle in a laboratory setting, e.g. in Mayhorn & Nyeste's work [14]. Because phishing emails can be considered a binary classification task [26], such designs often rely on a combination of a mail sorting task with a binary forced choice response format as their measure of phishing detection [13]. All the mentioned designs are usually coupled with additional measurements of influence variables like demographics, psychological variables or contextual-situational cues.

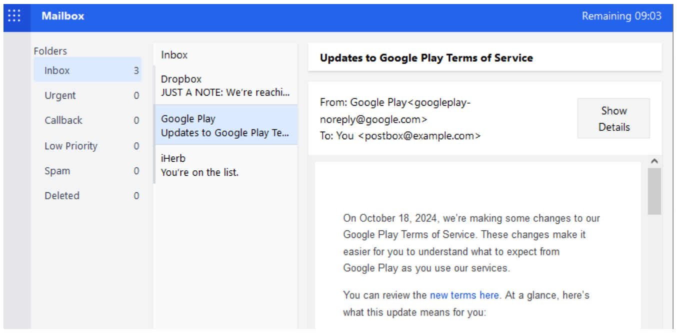
Research using such designs has found influence of personality variables like extraversion on phishing detection [10]; Cognitive reflection and sensation seeking were also identified as factors influencing phishing susceptibility [11]. Work by Parsons et al. found participants with higher cognitive impulsivity to be more susceptive to phishing, while work experience showed an effect towards better phishing detection [25]. Mayhorn & Nyeste found working memory capacity and irrelevant memory inhibition to be positively associated with phishing detection [14]. Seminal research by Downs et al. suggests that low familiarity with phishing raised susceptibility to phishing [24]. Regarding circumstances, scenario-based research suggests that knowledge about web structure and technology aids in phishing detection [27], [28]. Zheng & Becker found in their scenario-based study that enforced header presentation of emails did not improve phishing detection performance [29].

C. Limitations of Scenario-Based Designs

The main limitations of scenario-based studies are their generalizability and ecological validity, which stems from their design. While this is a general problem in Usable Security research [30], scenario-based designs exacerbate this issue as many studies tell their participants beforehand that the task they will be undertaking is about phishing. This will lead to effect misinterpretations due to higher alertness or vigilance [30], [31]. As has been demonstrated in other fields, scenario-based designs are highly sensitive to ecological validity problems, which for researchers often prove to be hard to impossible to solve [32]. Kieffer et al. define ecological validity as "representative users performing representative tasks in their natural environment" [33]. While the former two aspects, representative users and representative tasks, are often met, the latter is lacking in scenario-based designs: Most users will encounter phishing in their daily life, which differs substantially from a controlled laboratory setting. There will be external circumstances at play like distractions, or other primary tasks that demand attention [31]. As observational qualitative evidence suggests, such external influences are important to take into account [8] if research wants to be able to properly model influences on phishing detection.

III. Introducing PhishyMailbox

While scenario-based designs simulate the problem to gain insights, we deemed simulating the environment a more robust solution to collect natural behavioural data in a familiar and safe environment that adheres to ethical guidelines. We designed PhishyMailbox (available at https://github.com/Enterprize1/phishy-mailbox; Docker image provided at https://hub.docker.com/r/thorstenthiel/phishy-mailbox) for exactly this purpose. The typescript web-app simulates an email client akin to popular email front-ends in the browser (see Figure 1), combined with an admin interface for researchers or practitioners to administer, construct and design studies or tests.



Dropbox
JUST A NOTE: We're reachi...

Google Play
Updates to Google Play Te...

iHerb
You're on the list.

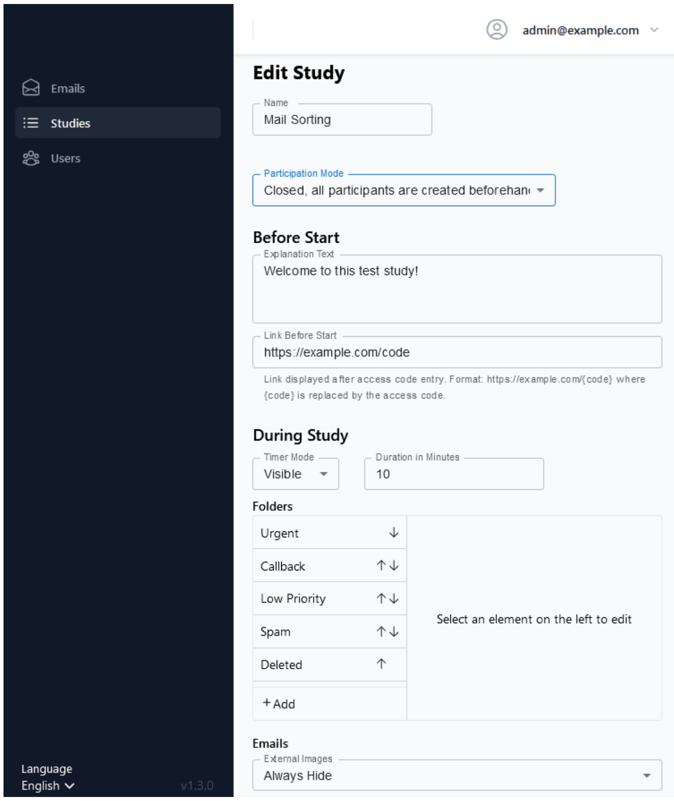
Updates to Google Play Terms of Service

From: Google Play<googleplay-noreply@google.com>
To: You <postbox@example.com>

On October 18, 2024, we're making some changes to our Google Play Terms of Service. These changes make it easier for you to understand what to expect from Google Play as you use our services.

You can review the new terms here. At a glance, here's what this update means for you:

Fig. 1: A browser screenshot of PhishyMailbox's participant user interface. It shows an exemplary inbox, running timer and emails to sort.



Edit Study

Name: Mail Sorting

Participation Mode: Closed, all participants are created beforehan

Before Start

Explanation Text: Welcome to this test study!

Link Before Start: https://example.com/code

Link displayed after access code entry. Format: https://example.com/{code} where {code} is replaced by the access code.

During Study

Timer Mode: Visible

Duration in Minutes: 10

Folders

Urgent
Callback
Low Priority
Spam
Deleted

Emails

External Images: Always Hide

Fig. 2: The researcher panel for PhishyMailbox, showing parts of the study configuration fields.

PhishyMailbox currently implements English and German user interfaces, which participants or researchers can switch between any time. The app is based on two components: Study participants interact with the Next.js front- and backend, which acts as a single-page app for the browser and provides an API to a PostgreSQL database, which stores all data.

A. Features For Researchers

PhishyMailbox is designed to offer flexibility for researchers and support common scientific workflows. The administrative interface can be protected with a custom username and password, multiple users are possible as well to manage access control. The interface allows uploading email files in EML format, editing them, add and edit studies (see Figure 2), as well as manage researcher user credentials. Uploaded emails as well as newly created ones can be edited at will: Header code and email body (HTML) can be freely designed; email sender, subject, and identifier are also subject to change. Creating or editing a study allows configuration of the following:

- Edit the welcome message as well as the outro message.
- Adding emails to studies and editing the study folder structure.
- Pre-define tokens for participants to log in or set a link to enable server hosting and remote participation — the latter will auto-generate tokens.
- Include links before and after the email task to seamlessly add surveys or other tasks to the study workflow. Participant tokens may be appended to any links, e.g. to pass them to survey platforms.
- Changing external image display options for participants.
- Set a timer and change timer display options for participants.

The advantage PhishyMailbox has over traditional methods of scenario-based research is that it by default collects activity data of study participants without having to specify it. The following interactions users show with the emails are logged as single data points, including timestamps:

- **Email view:** Email was opened.
- **Email details view:** The email header was viewed.
- **Email moved:** The email was moved into another folder. This action also saves the folder IDs (*from* and *to*).
- **Email scrolled:** The user scrolled in an email. The scroll position is also recorded.
- **Email link click:** The user clicks on a link in an email. The URL and the link text are recorded for the link.
- **Email link hover:** The user hovers over a link in an email with their cursor. The URL and the link text are recorded as well.
- **Start and end:** Records timestamps when users start and end the study.

All of this data is stored together with the participant token in the database and can be downloaded by researchers from the admin panel. Additionally, study structure, settings, and contents can be uploaded and downloaded as JSON files as well to ensure computational reproducibility.

B. Security and Ethics

Since this app was designed for phishing research, several safety-features were built into the app: An event listener was integrated to relay users to an ”Error 404” webpage after clicking links. Further, we used `<iframe>` HTML elements to isolate email content from the app itself. Using the sandbox parameter, we prevented navigation of the web-app as well as JavaScript execution from within an iframe. We applied the content security policy [34] to prevent loading from or navigating to external sources while being able to execute and analyse CSS rules. All of this enables the use of real phishing emails alongside legitimate ones while not disrupting the natural workflow of viewing and handling emails within a browser. Security of the executing machine can also not easily be compromised this way. Email attachments like PDF or office files are not accessible for the viewer, rendering an accidental inclusion of malicious attachments harmless.

To minimize ethical implications of PhishyMailbox usage, we designed the app so that it does not collect personally identifying information and isn't able to harm IT systems. Thus, we don't expect ethical issues to arise from the use of PhishyMailbox by itself. Should, however, other paradigms be coupled with PhishyMailbox, e.g. eye tracking applications, the ethical implications are bound to change. This requires further evaluation on the side of the researchers conducting the study, and, if need be, IRB approval.

C. *Limitations of PhishyMailbox*

Although we detailed the improvements of PhishyMailbox over previous approaches, the tool itself is not without limitations. Firstly, PhishyMailbox does not allow interaction with websites behind links, which limits its functionality in phishing research beyond the emails. This was a limitation we accepted since we wanted to prioritize security of the app to enable use of real phishing emails. Secondly, the user interface is not customizable by researchers or users. We decided against this to make research comparable within the PhishyMailbox paradigm. Lastly, the main limitation of PhishyMailbox is that it represents an emulation of a webmail client. As such, it does not provide the rich interaction possibilities such clients provide. Unfortunately, this is a trade-off effect, which exists because we wanted to enable use of real phishing emails while providing a secure environment. Researchers should consider these limitations before using the application.

IV. Usability evaluation

We evaluated PhishyMailbox usability to test the aptitude of the system in practice with a preliminary sample of researchers.

A. *Method*

We invited usable security and privacy scientists from our professional network to participate in a usability evaluation of PhishyMailbox. As our target sampling population was small and hard to reach, we could not meaningfully screen for specific criteria, as availability constraints of possible participants already rendered recruiting a difficult endeavour. Our preliminary sample consisted of 5 researchers with ages between 26 and 32 years (Mean: 28.6; SD: 2.70). They were familiar with phishing research, but had not published phishing papers of their own. Of our participants, 2 identified as female and 3 as male. The participants were previously informed that beyond familiarization with the software and future use if desired, there would be no reward for participation except for snacks and drinks, which we provided. We adhered to the ethical standards of the German Psychological Association for research [35], which the lead author of this work is part of: We firstly explained the purpose and broad makeup of the software as well as the upcoming evaluation to the participants, and obtained informed consent for participation. We then put forward a series of six tasks which the participants were asked to carry out using the software. We chose these tasks so that we could 1) enable participants to experience the main features of the software, 2) gain feedback on clarity and usability of the processes, and 3) familiarize the participants with the software and suggest improvements. The tasks included:

1) Logging in with provided login data.
2) Uploading new emails from a desktop folder.
3) Editing of one email by removing a symbol in the subject.
4) Creating a new study with previously uploaded emails.
5) Adding participants to the study manually.
6) Exporting of study data.

Participants were encouraged to give feedback, try to circumvent settings or security, and ask questions if necessary as they completed the task. We recorded all questions or suggestions directed at us during the task, as well as notable user interactions that were not part of the task. After the tasks, participants were asked to fill out demographic questions (gender and age) as well as the System Usability Scale (SUS) [36] regarding PhishyMailbox. We altered the first SUS question — "I think that I would like to use this system frequently" — to better fit the research context. Our altered version asked: "I think that I would like to frequently use this system for studies of this kind". After completion of the SUS, participants were thanked and given a link to the GitHub repository, as well as final contact details for questions regarding future use of the software.

B. *Qualitative Results*

Overall, participants expressed no concerns or suggestions regarding task 1 and 6 of the study. A few minor software bugs related to participant addition and uploading study JSON data were found, which we fixed prior to submission of this work.

The email upload (task 2) prompted 3 participants to ask whether multiple simultaneous uploads of emails were possible. As this was indeed possible at the time of evaluation, we took this as feedback that the description of the upload panel was in need of an update to reflect this possibility. Two participants suggested that the email editing menu in task 3 could benefit from a WYSIWYG editor in addition to the existing HTML editor. Additionally, one participant described the HTML editor as lacking line numbers and having inadequate scrolling options. Study creation (task 4) prompted a few questions from 3 participants about the concrete functionality of different data entry fields like the email selection field, the folder creation field and the participant addition. 2 participants asked about the specific details regarding generation of participant IDs for studies in task 5.

Based on these results, in a future version of PhishyMailbox, we will add tooltips to explain how participant IDs are generated and exported and to describe all fields in the study creation menu, clarifying functionality. The English data entry field descriptions will be updated for clarity, we will optimize the HTML editor and add a WYSIWYG editor as well. Additionally, we will compose a researcher use tutorial, which we will publish on GitHub alongside the software source code and documentation.

C. Quantitative Results

PhishyMailbox received SUS scores from 72.5 to 97.5, with an average of 85 points. This places the software in the top 25% of software regarding SUS usability evaluations [37]. Participants especially seemed to value the ease of access and simplicity, with questions 4,7, and 10 scoring very highly. A rather high amount of neutral ratings for question 8 (difficulty of use) likely indicates a lack of tooltips. These quantitative results are in line with our qualitative observations and will be addressed in future software versions as described in the previous section. As our sample size is still low, results cannot be considered representative, but this will be addressed with continued assessment with the goal of including researchers from other fields as well.

V. Future Work Opportunities with PhishyMailbox

PhishyMailbox integrates well into existing analysis strategies of phishing detection data and generates possibilities for new methods and assessment.

A. Applying Signal Detection Theory

Signal detection theory (SDT) [38], originating from psychophysics, is a method to analytically separate signal from noise. Applied to binary decision data in a test setting, it can be used to differentiate between response bias — a general tendency of the tested person to answer in a certain way — and sensitivity, the true ability of the person. SDT has been applied to phishing detection data of vignette-based experiments for almost a decade already [9], [12]. PhishyMailbox enables the use of SDT as well, if researchers employ a mail-sorting task for their study. While researchers have to specify what constitutes a hit, miss, false alarm or correct rejection by the participant, the range of data recorded by PhishyMailbox enables a plethora of possible feature combinations to define highly precise rules. Such designs are especially apt to research a broad range of influence factors on phishing detection from psychological to environmental, utilizing analytical frameworks like structural equation modelling.

B. Item Response Theory for Phishing Test Development

Item response theory (IRT) and its methods, e.g. the Rasch Model [39] or 2PL-Model [40], present a currently underutilized framework for phishing researchers. Since phishing detection in essence is a binary-outcome task, these models can be used to construct or evaluate phishing detection ability tests comprised of a mix of legitimate and phishing emails. Such an approach to date has not been published, probably because IRT has not yet permeated the Usable Security community, and such evaluations require high sample sizes that were hard to attain with previous experimental designs. The potential gains from a phishing detection test, for example measuring detection ability with PhishyMailbox in an assessment centre, can, however, not be understated: Knowing how well individuals can detect phishing could enable self-hosted skill-based training at overall low cost for organizations, while simultaneously improving employee security capabilities and fulfilling compliance requirements.

C. Eye Tracking Integration

At the time of writing, eye tracking has proven to be a useful tool to causally link user perceptions to their actions on screens for almost half a decade [41]. However, until recently, eye-tracking has been an expensive task, requiring special hardware and software, as well as a dedicated lab setup, considerably reducing study sample sizes of studies. Using eye tracking data to identify which exact email elements users are looking at when dealing with phishing emails has been done, albeit with small participant counts, limiting generalizability of such data [42], [43]. Recent research, however, has found the use of eye tracking with webcams [44], or even combining it with machine learning on mobile devices [45] a viable option. Abdrabou et al. in a preprint have successfully linked eye-tracking data with a mail sorting task, albeit on a small scale [46]. Linking such eye tracking data to PhishyMailbox presents an opportunity to analyse email decision-making and email perception on a large scale, enabling population-level inference about phishing detection. It seems feasible to write add-ons for PhishyMailbox integrating these approaches, effectively abolishing the lab-only limitation of most eye tracking study designs.

VI. Conclusion

Scenario-based experimental designs used in the phishing detection literature so far show considerable drawbacks when it comes to ecological assessment of user interactions with emails. To remedy this, we developed PhishyMailbox, a free and open source web app that provides a highly flexible, usable environment for researchers and practitioners interested in research or assessment of (phishing) emails. We evaluated the usability of PhishyMailbox with a small preliminary sample of usable security and privacy researchers. Qualitative insights gained from this evaluation shed light on necessary improvements regarding labelling and editor options, which will be addressed with continued app development. Our quantitative results demonstrate good overall usability of the app. The application shows great flexibility and can provide an excellent platform for future studies combining phishing detection research with item response theory, signal detection theory and (mobile) eye tracking approaches. We also welcome and encourage members of the research community to submit pull requests and future development ideas to foster phishing detection research and methodological rigour in usable security.

References

[1] M. Ell and S. Rizvi, "Cyber security breaches survey 2024," UK Department for Science, Innovation & Technology, Tech. Rep., 2024.

[2] D. Lain, K. Kostiainen, and S. Čapkun, "Phishing in Organizations: Findings from a Large-Scale and Long-Term Study," in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022. DOI: 10.1109/SP46214.2022.9833766.

[3] M. Volkamer, M. A. Sasse, and F. Boehm, "Analysing Simulated Phishing Campaigns for Staff," in *Computer Security*, I. Boureanu *et al.*, Eds., 2020, ISBN: 978-3-030-66504-3. DOI: 10.1007/978-3-030-66504-3_19.

[4] L. Brunken, A. Buckmann, J. Hielscher, and M. A. Sasse, "To Do This Properly, You Need More Resources: The Hidden Costs of Introducing Simulated Phishing Campaigns," 2023, ISBN: 978-1-939133-37-3.

[5] M. A. Sasse, J. Hielscher, J. Friedauer, and A. Buckmann, "Rebooting IT Security Awareness – How Organisations Can Encourage and Sustain Secure Behaviours," in *Computer Security. ESORICS 2022 International Workshops*, S. Katsikas *et al.*, Eds., 2023, ISBN: 978-3-031-25460-4. DOI: 10.1007/978-3-031-25460-4_14.

[6] Q. Cui, G.-V. Jourdan, G. V. Bochmann, I.-V. Onut, and J. Flood, "Phishing Attacks Modifications and Evolutions," in *Computer Security*, J. Lopez, J. Zhou, and M. Soriano, Eds., 2018, ISBN: 978-3-319-99073-6. DOI: 10.1007/978-3-319-99073-6_12.

[7] F. Carroll, J. A. Adejobi, and R. Montasari, "How Good Are We at Detecting a Phishing Attack? Investigating the Evolving Phishing Attack Email and Why It Continues to Successfully Deceive Society," *SN Computer Science*, 2022. DOI: 10.1007/s42979-022-01069-1.

[8] V. Distler, "The Influence of Context on Response to Spear-Phishing Attacks: An In-Situ Deception Study," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3581170.

[9] C. I. Canfield, B. Fischhoff, and A. Davis, "Quantifying Phishing Susceptibility for Detection and Behavior Decisions," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2016. DOI: 10.1177/0018720816665025.

[10] P. Lawson, C. J. Pearson, A. Crowson, and C. B. Mayhorn, "Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy," *Applied Ergonomics*, 2020. DOI: 10.1016/j.apergo.2020.103084.

[11] H. S. Jones, J. N. Towse, N. Race, and T. Harrison, "Email fraud: The search for psychological predictors of susceptibility," *PLOS ONE*, 2019. DOI: 10.1371/journal.pone.0209684.

[12] J. Martin, C. Dubé, and M. D. Coovert, "Signal Detection Theory (SDT) Is Effective for Modeling User Behavior Toward Phishing and Spear-Phishing Attacks," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2018. DOI: 10.1177/0018720818789818.

[13] D. Sturman, E. A. Bell, J. C. Auton, G. R. Breakey, and M. W. Wiggins, "The roles of phishing knowledge, cue utilization, and decision styles in phishing email detection," *Applied Ergonomics*, 2024. DOI: 10.1016/j.apergo.2024.104309.

[14] C. B. Mayhorn and P. G. Nyeste, "Training users to counteract phishing," *Work*, 2012. DOI: 10.3233/WOR-2012-1054-3549.

[15] S. M. Bellovin, "Spamming, phishing, authentication, and privacy," *Communications of the ACM*, no. 12, 2004. DOI: 10.1145/1035134.1035159.

[16] A. J. Ferguson, "Fostering E-Mail Security Awareness: The West Point Carronade," *Educause Quarterly*, 2005.

[17] M. Jakobsson and J. Ratkiewicz, "Designing ethical phishing experiments: A study of (ROT13) rOnl query features," in *Proceedings of the 15th international conference on World Wide Web*, 2006, ISBN: 978-1-59593-323-2. DOI: 10.1145/1135777.1135853.

[18] R. C. Dodge, C. Carver, and A. J. Ferguson, "Phishing for user security awareness," *Computers & Security*, 2007. DOI: 10.1016/j.cose.2006.10.009.

[19] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Communications of the ACM*, 2007. DOI: 10.1145/1290958.1290968.

[20] J. G. Mohebzada, A. E. Zarka, A. H. Bhojani, and A. Darwish, "Phishing in a university community: Two large scale phishing experiments," in *2012 International Conference on Innovations in Information Technology (IIT)*, 2012. DOI: 10.1109/INNOVATIONS.2012.6207742.

[21] K. Jansson and R. Von Solms, "Phishing for phishing awareness," *Behaviour & Information Technology*, 2013. DOI: 10.1080/0144929X.2011.632650.

[22] P. Kumaraguru *et al.*, "School of phish: A real-world evaluation of anti-phishing training," in *Proceedings of the 5th Symposium on Usable Privacy and Security*, 2009, ISBN: 978-1-60558-736-3. DOI: 10.1145/1572532.1572536.

[23] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Teaching Johnny not to fall for phish," *ACM Trans. Internet Technol.*, 2010. DOI: 10.1145/1754393.1754396.

[24] J. S. Downs, M. B. Holbrook, and L. F. Cranor, "Decision strategies and susceptibility to phishing," in *Proceedings of the second symposium on Usable privacy and security*, 2006, ISBN: 978-1-59593-448-2. DOI: 10.1145/1143120.1143131.

[25] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram, "Phishing for the Truth: A Scenario-Based Experiment of Users' Behavioural Response to Emails," in *Security and Privacy Protection in Information Processing Systems*, L. J. Janczewski, H. B. Wolfe, and S. Shenoi, Eds., 2013, ISBN: 978-3-642-39218-4. DOI: 10.1007/978-3-642-39218-4_27.

[26] K. Kaivanto, "The Effect of Decentralized Behavioral Decision Making on System-Level Risk," *Risk Analysis*, no. 12, 2014. DOI: 10.1111/risa.12219.

[27] M. Pattinson, C. Jerram, K. Parsons, A. McCormac, and M. Butavicius, "Why do some people manage phishing e-mails better than others?" *Information Management & Computer Security*, 2012. DOI: 10.1108/09685221211219173.

[28] J. S. Downs, M. Holbrook, and L. F. Cranor, "Behavioral response to phishing risk," in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, 2007, ISBN: 978-1-59593-939-5. DOI: 10.1145/1299015.1299019.

[29] S. Zheng and I. Becker, "Presenting suspicious details in User-Facing e-mail headers does not improve phishing detection," in *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, 2022, ISBN: 978-1-939133-30-4.

[30] V. Distler *et al.*, "A Systematic Literature Review of Empirical Methods and Risk Representation in Usable Privacy and Security Research," *ACM Trans. Comput.-Hum. Interact.*, 2021. DOI: 10.1145/3469845.

[31] S. Egelman, J. King, R. C. Miller, N. Ragouzis, and E. Shehan, "Security user studies: Methodologies and best practices," in *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, 2007, ISBN: 978-1-59593-642-4. DOI: 10.1145/1240866.1241089.

[32] J.-H. Kim and S. Jang, "A scenario-based experiment and a field study: A comparative examination for service failure and recovery," *International Journal of Hospitality Management*, 2014. DOI: 10.1016/j.ijhm.2014.05.004.

[33] S. Kieffer, U. B. Sangiorgi, and J. Vanderdonckt, "ECOVAL: A Framework for Increasing the Ecological Validity in Usabil-

ity Testing," in *2015 48th Hawaii International Conference on System Sciences*, 2015. DOI: 10.1109/HICSS.2015.61.

[34] S. Stamm, B. Sterne, and G. Markham, "Reining in the web with content security policy," in *Proceedings of the 19th international conference on World Wide Web*, 2010, ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772784.

[35] Deutsche Gesellschaft für Psychologie e.V., *Berufsethische Richtlinien des Berufsverbandes Deutscher Psychologinnen und Psychologen e. V. und der Deutschen Gesellschaft für Psychologie e. V.* 2022. [Online]. Available: https://www.dgps.de/fileadmin/user_upload/PDF/Berufsetische_Richtlinien/BER-Foederation-20230426-Web-1.pdf (visited on 09/15/2023).

[36] J. Brooke, "SUS: A 'Quick and Dirty' Usability Scale," in *Usability Evaluation In Industry*, P. W. Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester, Eds., 1st ed., 1996, ISBN: 978-0-42915-701-1.

[37] A. Bangor, P. Kortum, and J. Miller, "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale," *Journal of User Experience*, 2009.

[38] D. M. Green and J. A. Swets, *Signal detection theory and psychophysics*. John Wiley, 1966.

[39] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*. The Danish Institute of Educational Research, 1960.

[40] A. Birnbaum, "Some latent train models and their use in inferring an examinee's ability," in *Statistical theories of mental test scores.* F. M. Lord and M. R. Novick, Eds., 1968.

[41] C. Ware and H. H. Mikaelian, "An Evaluation of an Eye Tracker as a Device for Computer Input," in *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface*, 1986, ISBN: 978-0-89791-213-6. DOI: 10.1145/29933.275627.

[42] K. Pfeffel, P. Ulsamer, and N. H. Müller, "Where the User Does Look When Reading Phishing Mails – An Eye-Tracking Study," in *Learning and Collaboration Technologies. Designing Learning Experiences*, P. Zaphiris and A. Ioannou, Eds., 2019, ISBN: 978-3-030-21814-0. DOI: 10.1007/978-3-030-21814-0_21.

[43] J. McAlaney and P. J. Hills, "Understanding Phishing Email Processing and Perceived Trustworthiness Through Eye Tracking," *Frontiers in Psychology*, 2020. DOI: 10.3389/fpsyg.2020.01756.

[44] X. Yang and I. Krajbich, "Webcam-based online eye-tracking for behavioral research," *Judgment and Decision Making*, 2021. DOI: 10.1017/S1930297500008512.

[45] N. Valliappan *et al.*, "Accelerating eye movement research via accurate and affordable smartphone eye tracking," *Nature Communications*, 2020. DOI: 10.1038/s41467-020-18360-5.

[46] Y. Abdrabou *et al.*, *Revealing the hidden effects of phishing emails: An analysis of eye and mouse movements in email sorting tasks*, arXiv: 2305.17044 [cs.HC], 2023. [Online]. Available: https://arxiv.org/abs/2305.17044.