You Know What? - Evaluation of a Personalised Phishing Training Based on Users' Phishing Knowledge and Detection Skills

Lorin Schöni
ETH Zurich
Zurich, Switzerland
lorin.schoeni@gess.ethz.ch

Victor Carles
ETH Zurich
Zurich, Switzerland
victor.carles92@gmail.com

Martin Strohmeier
armasuisse
Zurich, Switzerland
martin.strohmeier@armasuisse.ch

Peter Mayer
University of Southern Denmark
Odense, Denmark
Karlsruhe Institute of Technology
Karlsruhe, Germany
mayer@imada.sdu.dk

Verena Zimmermann
ETH Zurich
Zurich, Switzerland
verena.zimmermann@gess.ethz.ch

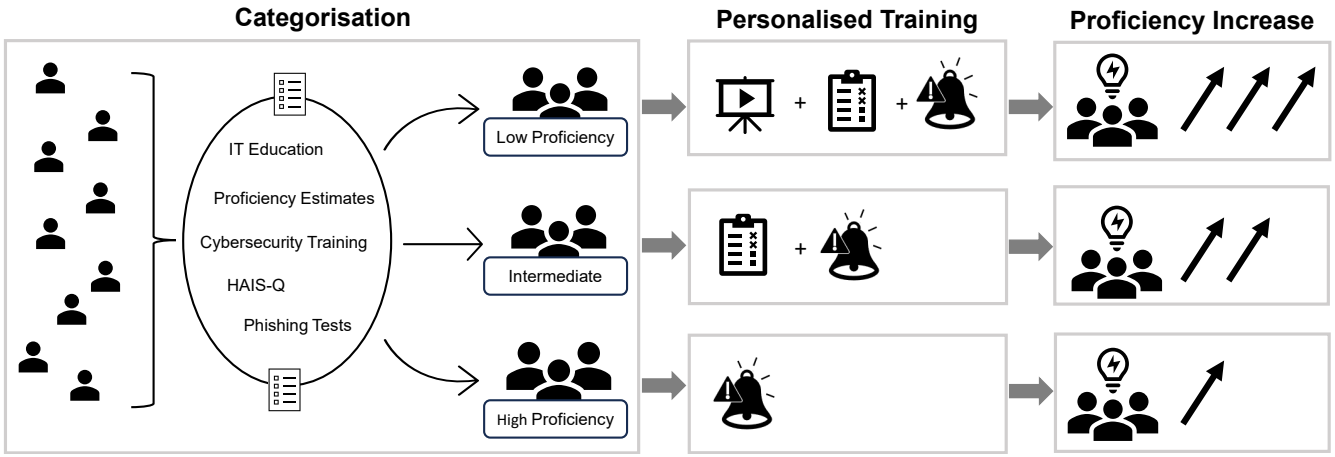


Figure 1: Participants are categorised based on knowledge, phishing detection ability, and other variables. The variables listed were the most helpful in differentiating between users. They then undergo a personalised training that consists of different combinations of training elements aimed at bringing all participants to a high proficiency level, i.e., the low proficiency group receives the highest number of training elements and shows the highest level of phishing proficiency improvements due to tailored training content.

ABSTRACT

Training is important to support users in phishing detection. To better match phishing training content with users' current skills, personalised training has huge potential. Therefore, we evaluated personalised training with N=96 participants in an online study. Participants were assigned to one of three groups based on a phishing proficiency score and received tailored training material. The training enhanced overall phishing proficiency, but also levelled the playing field, bringing all groups, regardless of their initial proficiency, to an equivalent post-training phishing proficiency level. For group assignment, the findings show that most person-related information, like age or personality traits, do not seem to affect phishing proficiency in a meaningful way. Yet, security awareness scales like the HAIS-Q or SA-13 seem to be useful indicators of phishing proficiency. The results demonstrate the feasibility of a personalised phishing intervention using relatively sparse data for categorisation into groups that receive tailored training content. Further research is needed to systematically evaluate the benefits and challenges of personalised phishing training.

EuroUSEC 2024, September 30–October 01, 2024, Karlstad, Sweden

ACM ISBN 979-8-4007-1796-3/24/09
https://doi.org/10.1145/3688459.3688460

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Security and privacy** → **Human and societal aspects of security and privacy**.

KEYWORDS

phishing detection, training, personalisation, phishing knowledge

ACM Reference Format:
Lorin Schöni, Victor Carles, Martin Strohmeier, Peter Mayer, and Verena Zimmermann. 2024. You Know What? - Evaluation of a Personalised Phishing Training Based on Users' Phishing Knowledge and Detection Skills. In *The 2024 European Symposium on Usable Security (EuroUSEC 2024), September 30–October 01, 2024, Karlstad, Sweden.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3688459.3688460

1 INTRODUCTION

Phishing is an acute cybersecurity challenge. The number of phishing emails is both at an all-time high and still expected to rise [4, 51]. To counter this threat, technical barriers are useful [52], but human-centred approaches including training are equally essential [51, 54]. Cybersecurity training usually serves specific goals, such as increasing knowledge [28], raising awareness [24], making users more alert [26], or nudging users [55]. Despite recognition of the human factor in cybersecurity [51, 54] and the importance of context [5, 12], most phishing training follows a 'one size fits all' approach (e.g., [10, 13, 28, 50]). Accordingly, they neglect individual differences [1, 11, 20, 48, 49]. Personalising phishing training based on user expectations, knowledge, experience, or context can be highly beneficial and increase its effectiveness [1, 25]. Personalised approaches have been proposed as a key factor for intervention success [5], are shown to enhance the effectiveness of anti-phishing training [25], and are recommended as central success factors by cybersecurity professionals [20]).

There are several barriers to the implementation of personalisation in cybersecurity training. For one, there is a lack of knowledge on *how* training should be personalised, i.e., *what data* should be collected, *which factors* should be considered, or *to what degree* the content is modified. Furthermore, institutions tend to focus on easy-to-implement training due to limited cybersecurity budgets [1], while also complying with data privacy standards [46]. Both factors might hinder the development or application of personalised training. Therefore, in this work, we evaluate a phishing training that can be personalised with comparably little effort to great effect and investigate to what degree privacy-invasive information needs to be factored into the personalisation.

Research Aim. The aim of this work was to explore the potential of personalised phishing training. In particular, we analysed whether presenting different training elements based on variables related to phishing proficiency (e.g., phishing knowledge or phishing detection capabilities) increases training effectiveness. Additionally, we wanted to explore the suitability of a single aggregate score as the basis for the classification.

We split this research aim into three concrete research questions: RQ1: What factors are the most helpful for differentiating between users' phishing proficiency and for selecting matching training components? RQ2: To what extent can a single composite score of various factors be used to categorise users into different proficiency levels? RQ3: To what extent can the education and awareness training elements for users with low or medium proficiency increase their proficiency to a similar level as high-proficiency users?

2 RELATED WORK

In learning research, personalisation has long been identified as a beneficial factor that substantially enhances learning [15], as it accommodates individual differences [11]. For instance, Klašnja-Milićević et al. [27] found that a learning software based on learning style and pre-existing knowledge substantially enhances test scores compared to a non-personalised control. A similar approach is used in adaptive learning, where training adapts in difficulty based on participant performance. Seda et al. [44] investigated adaptive learning in the cybersecurity context and found that it increased participants' training success rates. However, while adaptive learning techniques modify content based on user variables, they primarily do this by offering difficulty variations of the same content [23, 53]. Therefore, the learning experience is more accessible to low-performing users but does not address different needs or goals, such as by changing training elements or content.

Jampen et al. [25] conducted a comprehensive analysis of factors that affect phishing training effectiveness, finding that personalisation is a key factor as users differ in their capabilities. However, efforts in cybersecurity to provide personalised or interactive training forms (e.g., [10, 29, 45]) are hampered by limited resources [1]. Thus, static and generic material to educate users about phishing is still widely used [1] but suffers from low engagement [20]. Furthermore, existing approaches are rarely evaluated for population-specific characteristics and contexts, such as whether a certain user group benefits more from specific content.

Vasileiou and Furnell [48] discussed the mismatch between interpersonal differences and cybersecurity training. They highlighted the importance but also challenge of providing tailored education to account for these differences, such as knowledge or security awareness. Alotaibi et al. [2] proposed a personalised security awareness program framework. They outlined how user-specific factors, like prior knowledge or perception of security, are evaluated in a first step and then subsequently used to present modular training components. However, none of these personalisation approaches have yet been empirically evaluated in phishing training. The present study attempts to address this gap, by introducing and evaluating a personalised phishing training. To do so, we build on the security learning curve by Hielscher et al. [21] that integrates prior work from [8, 40]. As illustrated in Fig. 2, Sasse et al. [43] describe how the model postulates that a series of training steps is required to finally engage in and establish secure behaviours. First, users need to be informed about, sensitised, and knowledgeable about security. Thus, cybersecurity education is a first relevant step towards secure behaviours. On a higher level of the model, users need to develop self-efficacy in showing secure behaviours and practically implement their skills and abilities, i.e., a second relevant aspect is practical training. The highest levels of the model are concerned with embedding and habituating learned behaviours, e.g., through repetition or nudges, such as reminders. Reinheimer et al. [39] found that reminders such as interactive examples successfully enhanced phishing awareness after the effect of an initial training program had worn off. Based on the security learning curve [43] and the related findings on phishing reminders [39], our personalised phishing training approach includes three modules: (1) phishing awareness and (theoretical) education, (2) practical phishing training, and (3) phishing reminders.

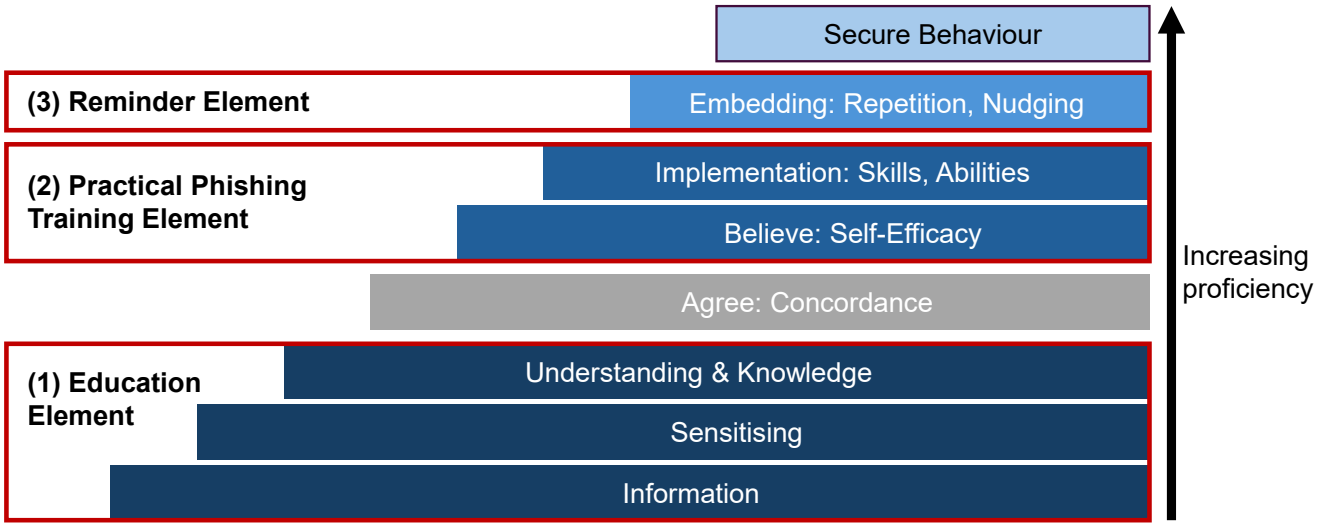


Figure 2: An illustration of the security learning curve adapted from [21] and described in [43]. The model describes a series of steps, building on each other, to ultimately reach secure behaviour. We superimpose the grouping of different steps into concrete training elements.

3 METHOD

To explore the effects of the personalised training intervention, we conducted an online between-subject study. Participants were assigned to one of the three training interventions through categorisation based on a phishing proficiency score. The proficiency score was calculated based on a pre-training questionnaire and evaluated again afterwards in a post-training questionnaire. To test the score classification, the functionality of the setup, and understandability of the questionnaire, we first conducted a pilot test with 10 participants followed by an additional pilot test with 20 participants after slightly adapting the score calculation. The following sections describe the sample, the score composition, the study procedure and material, as well as ethical considerations.

3.1 Participants

We initially recruited 120 participants from Prolific. We excluded 24 participants: 20 due to technical issues, three for automated or dubious responses, and one for missing data. The exclusions appeared random, with no notable differences among them. Thus, the final sample comprised N=96 participants. The participants' age distribution was as follows: 19 were between 18-24, 26 between 25-34, 23 between 35-44, 18 between 46-55, 7 between 56-65, and 3 between 66-75 years old. 63 participants had a university degree and 24 completed secondary education. 80 participants indicated they had some education background in IT, while 16 stated they did not. 54 participants stated they never completed a cybersecurity training before, while 26 participants completed training once, and 16 participants completed more than one training.

3.2 Personalisation and Categorisation

The training was personalised based on a phishing proficiency score, which was a combination of phishing-related variables from the two areas a) theoretical knowledge as, e.g., measured with a phishing knowledge quiz, and b) practical phishing detection capability, measured, e.g., through an email classification task and self-reported ability (see Appendix A). The proficiency score consisted of a combination of all variables and ranged from 21.95 to 49.2, with cut-offs for the low and high proficiency groups at 27.5 and 39.5, respectively. For example, for each out of five email screenshots that had to be classified as phishing or non-phishing, the user received two points for correct classification. For the detailed score calculation and assignment see Appendix B that also includes a hypothetical calculation example. Based on this score, participants were categorised into one of three different groups (see Fig. 3):

1) **Low proficiency:** In line with the security learning curve [43] people with low proficiency first need to learn what phishing is, why it is relevant, and how to detect it before they can successfully detect phishing emails. Hence, the group first received an educational video[1] targeting ***awareness and theoretical knowledge***. Then, the transfer of the theoretical knowledge to everyday life situations, i.e., the implementation of skills and abilities [43], needs to be practised. Therefore, the group additionally completed a ***practical phishing detection training*** (see Fig. 7 in Appendix C). And finally, in everyday life the alertness for phishing might decrease over time or in stressful situations. Repetition and nudges such as reminders can support successful habituation of the learned behaviour [43]. Hence, the group finally received ***reminders***, informed by the findings of [39] to keep alertness levels high.
2) **Intermediate proficiency:** People with intermediate proficiency know about phishing in theory but may lack practise.

[1]What is Phishing? - https://youtu.be/WG8V1_Sj5g0

Hence, they only completed the ***practical detection task*** and also received ***reminders*** to keep alertness levels high.

3) **High proficiency:** People with high proficiency know about phishing and can successfully detect phishing emails. However, even they might lack alertness in everyday life, preventing them from successfully applying their skills. Hence, this group only received ***reminders*** to counteract a lack of alertness.

3.3 Procedure

After providing informed consent, participants completed the pre-training questionnaire, including a theoretical knowledge and practical phishing detection test. The theoretical test consisted of multiple-choice questions evaluating participants' knowledge, where one out of multiple answers was correct, as well as selected items from the Human Aspects of Information Security questionnaire (HAIS-Q) [36] to capture attitudes and behaviour relating to email and internet use. One example is "It's risky to open an email attachment from an unknown sender." (see Appendix A.2 for a detailed overview). The multiple-choice questions shown on Appendix A.3 queried knowledge about phishing, such as asking people "What is phishing?" or "If you fall for a phishing scam, what should you do to limit the damage?" The order of these questions was randomised across the study, to account for potential differences in difficulty. In the practical detection test, we presented five emails to participants, prompting them to classify whether each is phishing or not. Additionally, we asked them to self-report their knowledge, ability to detect, and level of alertness on a scale from 1 to 5. Afterwards, participants were classified into the three proficiency level groups. The variables were weighted with the goal of differentiating between participants on knowledge and awareness levels, matching the additional training components that low and medium proficiency users would receive.

An overview of the procedure is provided in Fig. 3. After the categorisation, participants interacted with the training for 15 minutes. To increase realism and keep all participants occupied for the same amount of time despite different training components, they completed a background task during the training. This task involved interacting with emails in a fictional mailbox with various actions, such as reading an attachment or sending a reply (see Appendix C). The training differed based on the proficiency level participants were assigned to as described in section 3.2 and shown in Fig. 7. The reminders informing of phishing threats were integrated into the background task throughout the training time, whereas the 2-minute educational video and the 5-minute phishing detection training were timed to appear after interacting with the background task for one and five minutes, respectively.

After the training, participants completed a post-training questionnaire that included a second phishing proficiency task, demographic and background information, as well as self-report items (see Appendix A.2). To not introduce variance in terms of the difficulty levels of the pre- and post phishing proficiency tasks, all participants received the same questions in a randomised order (see Appendix A.3). Furthermore, we evaluated personality aspects using the short version of the Big Five Inventory with ten items (BFI-10) questionnaire [38], security awareness using the Security Attitude Inventory with 13 items (SA-13) questionnaire [14], and privacy concerns using the Internet Users' Information Privacy Concerns scale with eight items (IUIPC-8) [18] (see Appendix A.2 for more information on the scales).

3.4 Ethical Considerations

The study design followed established ethical guidelines for psychological research involving humans [3] and was approved by our institution's ethics commission. We minimised the potential for privacy invasion, e.g., by collecting age ranges instead of a concrete age. Participants were informed about the nature of the tasks and provided informed consent. Participation was voluntary and participants could abort the study and request the deletion of their data at any time without negative consequences. All participants received an equal payment in line with Prolific's suggestions for fair compensation of GBP 4.50 for their 30-minute participation.

4 RESULTS

Based on the score assignment process, 7 participants were in the low proficiency, 54 in the intermediate and 35 in the high proficiency group. To account for the variability in group sizes and the small number of participants in the low proficiency group, we only conducted robust non-parametric tests [9, 22] for statistic analysis. We used the Kruskal-Wallis Rank Sum Test when comparing groups, pairwise Wilcoxon signed-rank tests with Bonferroni adjustments [6] when evaluating changes within groups, and Fisher's exact test when comparing categorical variables. Even though we relied on validated scales such as SA-13, the IUIPC, or the HAIS-Q, we verified the reliability for our sample. When we calculated internal consistency as an established reliability measure, all measures indicated acceptable or better reliability metrics between .75 and .88 (see Appendix D.3 for details).

Wilcoxon signed-rank tests with Bonferroni adjustments confirmed that aggregate scores increased after completing the training for the low proficiency ($z = -2.37, p = .047$), intermediate ($z = -6.24, p < .001$), and high proficiency groups ($z = -3.41, p = .002$). Furthermore, a Kruskal-Wallis test confirmed the score gain differed significantly between groups ($H(2, 96) = 39.75, p < .001$). As

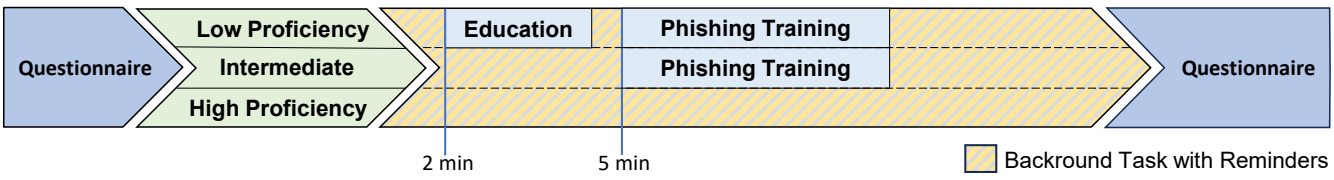


Figure 3: An overview of the categorisation and intervention procedure.

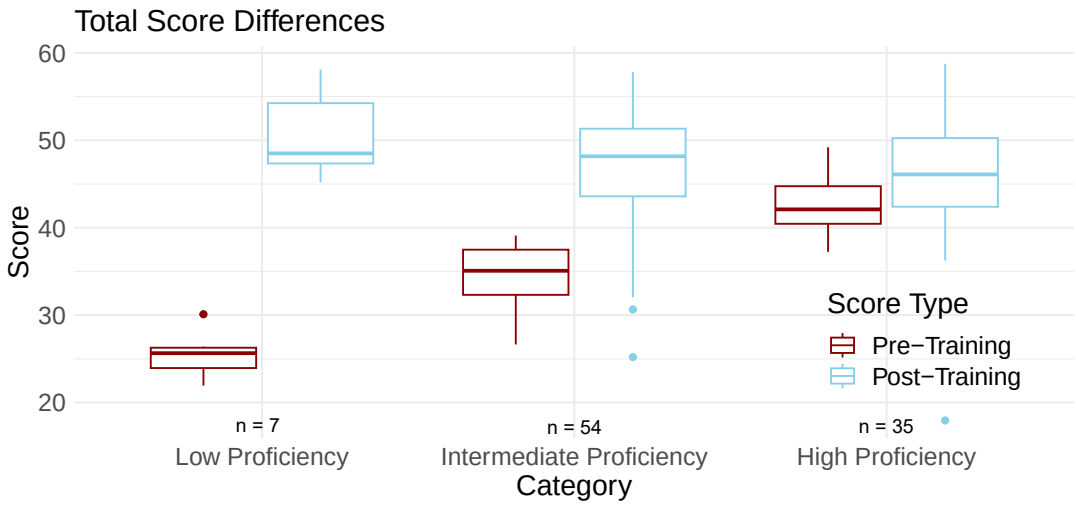


Figure 4: Comparisons of pre-training and post-training scores, separated by proficiency groups.

illustrated in Fig. 4, the low proficiency group improved the most, while the high proficiency group improved the least.

The mean phishing susceptibility, the probability of falling for a phishing email, decreased by 24%. In the pre-training classification task, A Kruskal-Wallis test revealed significant differences between groups ($H(2, 96) = 14.87, p < .001$). Mean susceptibility was highest in the low proficiency group with 34%, 31% in the intermediate, and 21% in the high proficiency group. However, in the post-training classification task, the groups' proficiency did not differ significantly any more ($H(2, 96) = 1.58, p = .454$).

We analysed how effective each score component in isolation was at differentiating between the groups, and whether other variables not used in the score calculation differed between the groups. Furthermore, we evaluated how the groups differed in other measures, such as overall phishing susceptibility, security awareness, or privacy concerns. Finally, we measured the effect of variables on score improvements.

4.1 Score Component Differences Between Groups

To understand how suitable each score component was, we first evaluated how effective they were at differentiating proficiency groups. The detailed results for each variable are shown in Table 2 in Appendix D.

Security Background & Self-Estimate: We measured whether security-related background information and proficiency self-estimates differed between groups. A Fisher's exact test confirmed that previous security training ($p = .001$) and IT education background ($p < .001$) significantly differed between proficiency groups. Kruskal-Wallis tests also confirmed that self-reported knowledge ($H(2, 94) = 9.80, p = .007$), ability ($H(2, 94) = 10.50, p = .005$), and alertness ($H(2, 94) = 8.56, p = .014$) differed significantly across groups. All other variables did not show a significant difference between proficiency groups.

Security Intentions & Behaviour: We evaluated whether HAIS-Q scores differed significantly between groups. Out of all the HAIS-Q sub scales, a Kruskal-Wallis test revealed that only email attitudes differed significantly ($H(2, 94) = 8.03, p = .018$).

4.2 Other Group Differences

Security Awareness: The SA-13 sub scales attentiveness ($H(2, 94) = 6.94, p = .031$) and resistance ($H(2, 94) = 7.40, p = .025$) differed significantly between groups, while engagement ($H(2, 94) = 4.07, p = .130$) and concernedness ($H(2, 94) = 1.51, p = .469$) did not significantly differ between groups. Overall, the total SA-13 score was significantly related ($H(2, 94) = 9.05, p = .011$) with a moderate effect ($\eta^2 = .076$), stronger than that of the individual sub scales attentiveness ($\eta^2 = .053$) and resistance ($\eta^2 = .058$).

Privacy Concern: We captured privacy concerns using the IUIPC-8 scale [18]. The overall score ($H(2, 94) = 7.55, p = .023$), the control ($H(2, 94) = 6.18, p = .045$) and the collection sub scale differed significantly between groups ($H(2, 94) = 6.51, p = .039$), while the awareness sub scale ($H(2, 94) = 2.27, p = .322$) did not.

Personality: We created linear regression models containing all personality traits for pre- and post-training scores (see Table 3 and Table 4 in Appendix D). We found no significant influence of any personality traits pre-training, while extraversion was the only significant influence on post-training scores ($\beta = -1.83, t(90) = -2.14, p = .035$), with which it correlated weakly ($r = -.15$).

4.3 Effect on Training Improvements

For all participants, we calculated the score difference before and after the training. We then used Pearson's product-moment correlation to assess the effect of the HAIS-Q on training improvements. The HAIS-Q values correlated weakly and negatively with the practical score ($r = -.13, t(94) = -2.68, p = .009$) indicating that people with higher pre-training HAIS-Q scores benefited less from the training, likely because the values were already higher to start with. This effect was even more pronounced ($r = -.27, t(94) = -2.73, p =$

.008) for the theoretical score gain, reflecting the close association with HAIS-Q and phishing knowledge and attitudes.

5 DISCUSSION

To summarise the key findings, the personalised training was effective in increasing participants' proficiency and reducing phishing susceptibility overall. The score gain differed between groups, with the low proficiency group seeing the highest increase and the high proficiency group the lowest. Phishing susceptibility decreased likewise, and group differences in proficiency seemed to disappear in the post-training classification performance. Overall, the findings indicate that the personalisation of components successfully enabled people with previously different proficiency levels to reach similar proficiency levels after the training. In practical terms, a benefit for people with already medium or high proficiency levels might be that they do not need to go through aspects they are already proficient in and can be spared effort and hence potential frustration. Because institutions' training budgets are often small and employees are assigned little time to complete cybersecurity training [1], personalisation has huge potential to keep training short while still providing effective results.

However, the findings should be interpreted cautiously given the exploratory score calculation and categorisation process and the large differences in group sizes. Yet, this study provides a first step towards empirically evaluating the effects of personalised phishing training and of the factors that influence proficiency levels and might thus be more or less relevant for future personalised approaches. For example, our findings indicated none or weak relationships with personality traits whereas security awareness scales like the HAIS-Q or SA-13 seem to be relevant indicators of phishing proficiency. In the following, we discuss these findings in further detail and in relation to previous work.

5.1 RQ1: Impactful Variables and Group Differences

In order to answer RQ1 related to differentiating user groups, we discuss variables and how they differ between groups below. Out of questions concerning the personal background, only previous security training, self-reported phishing proficiency, and IT education background differed significantly between groups. These data are relatively easy to collect and seem to influence user's security behaviour and knowledge greatly. Yet, other demographic information like age or education were not found to be significantly related to phishing proficiency. This contradicts many earlier findings (e.g., [45]), but is consistent with newer insights showing a more nuanced web of influences [41, 42] and studies that show these correlations simply disappear when controlling for confounding factors (e.g., [30]).

Personality. We also evaluated personality, as previous studies suggested various influences of personality traits on phishing susceptibility [17, 31]. However, we only found a minor difference of extraversion between the groups, suggesting that personality traits are not important to consider for categorisation.

Privacy Concerns. Many variables that might cause privacy-related concerns, such as personality, age, or email usage behaviour do not seem influential in affecting user's phishing proficiency, nor moderate training effects. Thus, the use of collecting these potentially privacy-invasive data might might not be justified for phishing training. Instead, user training could be personalised to provide suitable while still preserving user's anonymity and complying with regulations. This is especially important, as compliance and data privacy are a high priority for CISOs [46].

Efficacy of Categorisation. The HAIS-Q items used as a component in calculating participant's score seemed to correlate with participant's score overall. Unsurprisingly, these items correlated higher with the theoretical component as compared to the practical score component, as the HAIS-Q is based on hypothetical scenarios and principles based on the Knowledge–Attitude–Behaviour model, and does not directly test practical abilities [36]. However, only email and internet usage attitudes differed significantly between the groups. The overall SA-13 score differed between the groups with a moderate effect. Due to the economical nature of the SA-13, it, or perhaps even the shorter SA-6 variant, might serve as a good component of any future categorisation.

5.2 RQ2: Categorisation and Training Material.

Regarding the use of a single composite score posited in RQ2, we found that using a single score is possible to differentiate groups sufficiently. However, the use of score sub components, e.g. a theoretical and a practical score as in our case, appears to be helpful to distinguish user proficiency on different levels of the learning curve. Thus, for future work we suggest a modular score to not only distinguish different levels of the learning curve but also perhaps different sets of competencies.

Categorisation Procedure. We categorised participants into proficiency levels as many have moderate or high proficiency and benefit from distinct training, as based on [43], while fewer have low proficiency and need foundational knowledge. This allowed us to demonstrate clear benefits from separating participants into groups. Still, further work could explore other compositions or create more granular training steps. The findings also demonstrated the feasibility of using composite scores for categorisation. The three proficiency groups differed on self-estimated proficiency, phishing susceptibility, previous security training, and IT education background, as well as security awareness and intention indicators. While the difficulty in balancing group sizes highlights challenges of the categorisation, the size differences do not necessarily imply an unrealistic representation. Previous studies have consistently found a small percentage of individuals who exhibit especially high phishing susceptibility [47].

5.3 RQ3: Training Effects and Material

To address RQ3 and the extent to which our modular training elements were able to enhance phishing proficiency, we first discuss overall effects, and then the training material in more detail.

Phishing Proficiency Improvements. We were able to demonstrate that education and awareness training can significantly narrow the proficiency gap between low or medium proficiency users and their high-proficiency counterparts. While high proficiency users may not reach identical levels, the training effectively levels the playing field. This is especially promising, as a small percentage

of low-proficiency user has consistently been identified in previous research [30, 47]. Training that specifically targets these users may be a promising approach to lowering an organisation's overall vulnerability to phishing.

Choice of Educational Material. The short video-based material we selected for the low proficiency group has been chosen as a supposedly more engaging education element [7, 39] as compared to textual education material. Yet, all participants in the low proficiency group saw the same video. For future work, providing different materials to users based on their needs seems promising. This means personalising not only the composition of training elements but also the materials therein. Such an approach could improve willingness to engage with the material, as users have displayed a lack of enthusiasm and boredom in other contexts where training material is too generic and does not match user expectations and needs [20]. This could further be applied to different roles within an institution, which differ in their tasks and are targeted by different types of phishing attacks [1].

5.4 Limitations and Future Work

The study was exploratory in nature, providing a first step towards evaluating the effectiveness of presenting different training components to user groups based on phishing proficiency levels. Accordingly, the categorisation process had not been validated before but incorporated insights from previous research as a starting point. Furthermore, the training and data collection took place as part of a single online session, with no long-term measurement; an issue that is affecting many phishing studies [16, 30]. Therefore, we cannot evaluate the stability of these effects over time.

The large variability in group size skews the results and needs to be kept in mind when interpreting the findings. This variability was largely a consequence of using cut-offs defined a priori, which did not result in equally distributed group sizes despite pilot testing. Future work might be able to better account for this aspect by adjusting our exploratory score calculation informed by relevant factors extracted from related work.

Finally, the additive composition of training components across groups allowed for between-group comparisons. As such, we could initially demonstrate the benefits of personalised training. Yet, for future work it would be beneficial to additionally compare the personalised assignment to a group for which the training assignment is randomised, i.e., not influenced by the proficiency score. This would enable us to directly compare the potential benefit of personalisation as a whole to a non-personalised alternative to further validate the potential of personalisation in phishing training. Furthermore, future work could evaluate matching of training content and proficiency levels in more isolated ways, e.g., by providing only education content to high-proficiency users. Such a comparison could more directly evaluate the effectiveness of isolated components for specific user groups.

6 CONCLUSION

We evaluated a personalised training on N=96 participants, finding that it increased overall phishing proficiency, but showed higher increases for participants in lower pre-training proficiency groups. As participants with varying initial proficiency levels reached similar post-training proficiency levels, the personalisation of training seems promising in bridging proficiency gaps. Our study also demonstrated the feasibility of using a composite phishing proficiency score to categorise participants based on proficiency levels, and identified factors that differentiated well between the groups, without needing to rely on potentially privacy-invasive information like personality or age. While our study provides an initial empirical evaluation of personalised phishing training, further research is essential to investigate longitudinal effects, the assignment mechanism, and the training content.

7 DATA AVAILABILITY STATEMENT

The data that support the findings of this article are openly available in https://doi.org/10.3929/ethz-b-000689288.

ACKNOWLEDGMENTS

This work was graciously supported by armasuisse Science and Technology.

REFERENCES

[1] Hussain Aldawood and Geoffrey Skinner. 2019. Reviewing Cyber Security Social Engineering Training and Awareness Programs—Pitfalls and Ongoing Issues. *Future Internet* 11, 3 (March 2019), 73. https://doi.org/10.3390/fi11030073

[2] S. Alotaibi, Steven Furnell, and Y. He. 2023. Towards a Framework for the Personalization of Cybersecurity Awareness. In *Human Aspects of Information Security and Assurance (IFIP Advances in Information and Communication Technology)*, Steven Furnell and Nathan Clarke (Eds.). Springer Nature Switzerland, Cham, Switzerland, 143–153. https://doi.org/10.1007/978-3-031-38530-8_12

[3] APA. 2017. *Ethical principles of psychologists and code of conduct.* https://www.apa.org/ethics/code

[4] APWG. 2023. Phishing Activity Trends Report, 1st Quarter 2023. https://docs.apwg.org/reports/apwg_trends_report_q1_2023.pdf

[5] Steffen Bartsch and Melanie Volkamer. 2012. Towards the Systematic Development of Contextualized Security Interventions. In *Designing Interactive Secure Systems: Workshop at British HCI 2012, University of Birmingham, 11th September 2012.* BCS Learning & Development. https://doi.org/10.14236/ewic/HCI2012.69

[6] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[7] Benjamin M. Berens, Mattia Mossano, and Melanie Volkamer. 2024. Taking 5 minutes protects you for 5 months: Evaluating an anti-phishing awareness video. *Computers & Security* 137 (Feb. 2024), 103620. https://doi.org/10.1016/j.cose.2023.103620

[8] Marcus Beyer, Sarah Ahmed, Katja Doerlemann, Simon Arnell, Simon Parkin, Angela Sasse, and Neil Passingham. 2016. Awareness is only the first step: A framework for progressive engagement of staff in cyber security.

[9] Patrick D Bridge and Shlomo S Sawilowsky. 1999. Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t-test and Wilcoxon Rank-Sum Test in Small Samples Applied Research. *Journal of Clinical Epidemiology* 52, 3 (1999), 229–235. https://doi.org/10.1016/S0895-4356(98)00168-1

[10] Deanna D. Caputo, Shari Lawrence Pfleeger, Jesse D. Freeman, and M. Eric Johnson. 2014. Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy* 12, 1 (2014), 28–38. https://doi.org/10.1109/MSP.2013.106

[11] Sherry Y. Chen and Jen-Han Wang. 2021. Individual differences and personalized learning: a review and appraisal. *Universal Access in the Information Society* 20, 4 (Nov. 2021), 833–849. https://doi.org/10.1007/s10209-020-00753-4

[12] Nabin Chowdhury and Vasileios Gkioulos. 2021. Cyber security training for critical infrastructure protection: A literature review. *Computer Science Review* 40 (2021), 100361. https://doi.org/10.1016/j.cosrev.2021.100361

[13] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence Italy). ACM, New York, NY, USA, 1065–1074. https://doi.org/10.1145/1357054.1357219

[14] Cori Faklaris, Laura Dabbish, and Jason I. Hong. 2022. Do They Accept or Resist Cybersecurity Measures? Development and Validation of the 13-Item Security

Attitude Inventory (SA-13). https://doi.org/10.48550/arXiv.2204.03114
[15] Rida Indah Fariani, Kasiyah Junus, and Harry Budi Santoso. 2023. A Systematic Literature Review on Personalised Learning in the Higher Education Context. *Technology, Knowledge and Learning* 28, 2 (June 2023), 449–476. https://doi.org/10.1007/s10758-022-09628-4
[16] Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. 2021. SoK: Still Plenty of Phish in the Sea - A Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research. USENIX Association, Berkeley, CA, USA, 358. https://www.usenix.org/conference/soups2021/presentation/franz
[17] Edwin Donald Frauenstein and Stephen Flowerday. 2020. Susceptibility to phishing on social network sites: A personality information processing model. *Computers & Security* 94 (2020), 101862. https://doi.org/10.1016/j.cose.2020.101862
[18] Thomas Groß. 2021-04-01. Validity and Reliability of the Scale Internet Users' Information Privacy Concerns (IUIPC). *Proceedings on Privacy Enhancing Technologies* 2021, 2 (2021-04-01), 235–258. https://doi.org/10.2478/popets-2021-0026
[19] Andrew F. Hayes and Jacob J. Coutts. 2020. Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But.... *Communication Methods and Measures* 14, 1 (Jan. 2020), 1–24. https://doi.org/10.1080/19312458.2020.1718629
[20] Wu He and Zuopeng Zhang. 2019. Enterprise cybersecurity training and awareness programs: Recommendations for success. *Journal of Organizational Computing and Electronic Commerce* 29, 4 (2019), 249–257. https://doi.org/10.1080/10919392.2019.1611528
[21] Jonas Hielscher, Annette Kluge, Uta Menges, and M. Angela Sasse. 2022. "Taking out the trash": Why security behavior change requires intentional forgetting. In *Proceedings of the 2021 New Security Paradigms Workshop*. Association for Computing Machinery, 108–122. https://doi.org/10.1145/3498891.3498902
[22] Myles Hollander and Douglas A. Wolfe. 1973. *Nonparametric statistical methods*. Wiley New York, New York.
[23] Shiu-Li Huang and Jung-Hung Shiu. 2012. A User-Centric Adaptive Learning System for E-Learning 2.0. *Journal of Educational Technology & Society* 15, 3 (2012), 214–225. https://www.jstor.org/stable/jeductechsoci.15.3.214
[24] Norman Hänsch and Zinaida Benenson. 2014. Specifying IT Security Awareness. In *2014 25th International Workshop on Database and Expert Systems Applications*. 326–330. https://doi.org/10.1109/DEXA.2014.71
[25] Daniel Jampen, Gürkan Gür, Thomas Sutter, and Bernhard Tellenbach. 2020. Don't click: towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences* 10, 1 (2020), 33. https://doi.org/10.1186/s13673-020-00237-7
[26] Matthew L. Jensen, Michael Dinger, Ryan T. Wright, and Jason Bennett Thatcher. 2017. Training to Mitigate Phishing Attacks Using Mindfulness Techniques. *Journal of Management Information Systems* 34, 2 (April 2017), 597–626. https://doi.org/10.1080/07421222.2017.1334499
[27] Aleksandra Klašnja-Milićević, Boban Vesin, Mirjana Ivanović, and Zoran Budimac. 2011. E-Learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education* 56, 3 (April 2011), 885–899. https://doi.org/10.1016/j.compedu.2010.11.001
[28] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. 2009. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (New York, NY, USA) *(SOUPS '09)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/1572532.1572536
[29] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2010. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology* 10, 2 (2010), 1–31. https://doi.org/10.1145/1754393.1754396
[30] Daniele Lain, Kari Kostiainen, and Srdjan Capkun. 2022. Phishing in Organizations: Findings from a Large-Scale and Long-Term Study. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, New York, NY, USA, 842–859. https://doi.org/10.1109/sp46214.2022.9833766
[31] Pablo López-Aguilar and Agusti Solanas. 2021. Human Susceptibility to Phishing Attacks Based on Personality Traits: The Role of Neuroticism. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, New York, NY, USA, 1363–1368. https://doi.org/10.1109/COMPSAC51774.2021.00192
[32] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research* 15, 4 (2004), 336–355. https://doi.org/10.1287/isre.1040.0032 Place: US Publisher: Institute for Operations Research & the Management Sciences (INFORMS).
[33] Camila Paola Malkewitz, Philipp Schwall, Christian Meesters, and Jochen Hardt. 2023. Estimating reliability: A comparison of Cronbach's α, McDonald's ωt and the greatest lower bound. *Social Sciences & Humanities Open* 7, 1 (Jan. 2023), 100368. https://doi.org/10.1016/j.ssaho.2022.100368
[34] R. R. McCrae and O. P. John. 1992. An introduction to the five-factor model and its applications. *Journal of Personality* 60, 2 (June 1992), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x
[35] Alexis R. Neigel, Victoria L. Claypoole, Grace E. Waldfogle, Subrata Acharya, and Gabriella M. Hancock. 2020. Holistic cyber hygiene education: Accounting for the human factors. *Computers & Security* 92 (2020), 101731. https://doi.org/10.1016/j.cose.2020.101731
[36] Kathryn Parsons, Dragana Calic, Malcolm Pattinson, Marcus Butavicius, Agata McCormac, and Tara Zwaans. 2017. The Human Aspects of Information Security Questionnaire (HAIS-Q): Two further validation studies. *Computers & Security* 66 (2017), 40–51. https://doi.org/10.1016/j.cose.2017.01.004
[37] Robert A. Peterson. 1994. A Meta-analysis of Cronbach's Coefficient Alpha. *Journal of Consumer Research* 21, 2 (Sept. 1994), 381–391. https://doi.org/10.1086/209405
[38] Beatrice Rammstedt, Christoph J. Kemper, Mira Céline Klein, Constanze Beierlein, and Anastassiya Kovaleva. 2013. A Short Scale for Assessing the Big Five Dimensions of Personality: 10 Item Big Five Inventory (BFI-10). *methods, data, analyses* 7, 2 (2013), 17. https://doi.org/10.12758/mda.2013.013
[39] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, 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. Usenix, Berkeley, CA, USA, 259–284. https://www.usenix.org/conference/soups2020/presentation/reinheimer
[40] Karen Renaud, Melanie Volkamer, and Arne Renkema-Padmos. 2014. Why doesn't Jane protect her privacy?. In *Privacy Enhancing Technologies*, Emiliano De Cristofaro and Steven J. Murdoch (Eds.). Springer International Publishing, 244–262. https://doi.org/10.1007/978-3-319-08506-7_13
[41] Liliana Ribeiro, Inês Sousa Guedes, and Carla Sofia Cardoso. 2024. Which factors predict susceptibility to phishing? An empirical study. *Computers & Security* 136 (Jan. 2024), 103558. https://doi.org/10.1016/j.cose.2023.103558
[42] Dawn M. Sarno, Maggie W. Harris, and Jeffrey Black. 2023. Which phish is captured in the net? Understanding phishing susceptibility and individual differences. *Applied Cognitive Psychology* 37, 4 (2023), 789–803. https://doi.org/10.1002/acp.4075
[43] M. Angela Sasse, Jonas Hielscher, Jennifer Friedauer, and Annalina Buckmann. 2023. Rebooting IT security awareness–how organisations can encourage and sustain secure behaviours. In *Computer Security. ESORICS 2022 International Workshops*. Springer International Publishing, 248–265. https://doi.org/10.1007/978-3-031-25460-4_14
[44] Pavel Seda, Jan Vykopal, Valdemar Švábenský, and Pavel Čeleda. 2021. Reinforcing Cybersecurity Hands-on Training With Adaptive Learning. In *2021 IEEE Frontiers in Education Conference (FIE)*. IEEE, New York, NY, USA, 1–9. https://doi.org/10.1109/FIE49875.2021.9637252
[45] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010-04-10). Association for Computing Machinery, New York, NY, USA, 373–382. https://doi.org/10.1145/1753326.1753383
[46] Mario Silic and Andrea Back. 2016. The dark side of social networking sites: Understanding phishing risks. *Computers in Human Behavior* 60 (July 2016), 35–43. https://doi.org/10.1016/j.chb.2016.02.050
[47] Teodor Sommestad and Henrik Karlzén. 2019. A meta-analysis of field experiments on phishing susceptibility. In *2019 APWG Symposium on Electronic Crime Research (eCrime)*. 1–14. https://doi.org/10.1109/eCrime47957.2019.9037502
[48] Ismini Vasileiou and Steven Furnell. 2023. Enhancing Security Education - Recognising Threshold Concepts and Other Influencing Factors. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy - ICISSP*. SciTePress, 398–403. https://doi.org/10.5220/0006646203980403
[49] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H. Raghav Rao. 2011. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems* 51, 3 (2011), 576–586. https://doi.org/10.1016/j.dss.2011.03.002
[50] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. 2019. What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA) *(CHI '19)*. Association for Computing Machinery, 1–12. https://doi.org/10.1145/3290605.3300338
[51] Eva Wolfangel. 2023. The Human Element in Cybercrime and Cybersecurity. https://www.youtube.com/watch?v=LKUMRTLV49g
[52] Shouhuai Xu. 2019. Cybersecurity Dynamics: A Foundation for the Science of Cybersecurity. In *Proactive and Dynamic Network Defense*, Cliff Wang and Zhuo Lu (Eds.). Springer International Publishing, 1–31. https://doi.org/10.1007/978-3-030-10597-6_1
[53] Beste F. Yuksel, Kurt B. Oleson, Lane Harrison, Evan M. Peck, Daniel Afergan, Remco Chang, and Robert JK Jacob. 2016. Learn Piano with BACh: An Adaptive Learning Interface that Adjusts Task Difficulty Based on Brain State. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5372–5384. https://doi.org/10.1145/2858036.2858388
[54] Verena Zimmermann and Karen Renaud. 2019. Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. *International Journal of Human-Computer Studies* 131 (2019), 169–187. https://doi.org/10.1016/j.ijhcs.

2019.05.005
[55] Verena Zimmermann and Karen Renaud. 2021. The Nudge Puzzle: Matching Nudge Interventions to Cybersecurity Decisions. *ACM Transactions on Computer-Human Interaction* 28, 1 (Jan. 2021), 7:1–7:45. https://doi.org/10.1145/3429888

A QUESTIONNAIRES

A.1 Pre-Training Questionnaires

- Background
  - Have you ever taken part in a security training on phishing? [Never, Once, More than once]
  - How frequently do you use your mailbox? [once a week, once a day, multiple times a day, multiple times an hour]
  - How many emails do you receive on an average work day? [less than 5, less than 10, 10-50, more than 50]
  - How many spam or promotional emails do you receive on an average day? [less than 10, 10-50, more than 50, I don't know]
  - Do you use a professional mailbox in the context of your job? [yes, no, I don't have a job]
- Security Proficiency
  - Theoretical Phishing Test: 14 multiple choice questions.
  - Practical Phishing Test: Classification of 5 emails.
  - HAIS-Q [36]: The Human Aspects of Information Security Questionnaire measures the users' security dimensions knowledge, attitude, and behaviour across seven focus areas on a total of 63 items. The modular structure allows for selecting and focusing on specific focus areas (e.g. incident reporting) or dimensions (e.g., attitude). For our phishing questionnaire, we selected phishing-related items. Hence, the primary focus was on email use and internet use items, as they are most directly related to phishing. Furthermore, we focused on the dimensions attitude and behaviour as knowledge was already directly assessed through background information and in a knowledge quiz. We therefore used 6 items from email use sub scale [3 attitude, 3 behaviour] and 5 items from internet use sub scale [3 attitude, 2 behaviour] (one item excluded that concerned website safety as its phrasing does not directly relate to phishing):
    - Internet, Attitude: It can be risky to download files on my work computer. [scale from 1 - strongly disagree to 5 - strongly agree]
    - Internet, Attitude: Just because I can access a website at work, doesn't mean that it's safe. [scale from 1 - strongly disagree to 5 - strongly agree]
    - Internet, Attitude: If it helps me to do my job, it doesn't matter what information I put on a website. [scale from 1 - strongly disagree to 5 - strongly agree]
    - Internet, Behaviour: I download any files onto my work computer that will help me get my job done. [scale from 1 - strongly disagree to 5 - strongly agree]
    - Internet, Behaviour: When accessing the Internet at work, I visit any website that I want to. [scale from 1 - strongly disagree to 5 - strongly agree]
    - Email, Attitude: It's risky to open an email attachment from an unknown sender. [scale from 1 - strongly disagree to 5 - strongly agree]
    - Email, Attitude: It's always safe to click on links in emails from people I know. [scale from 1 - strongly disagree to 5 - strongly agree]
    - Email, Attitude: Nothing bad can happen if I click on a link in an email from and unknown sender. [scale from 1 - strongly disagree to 5 - strongly agree]
    - Email, Behaviour: I don't open email attachments if the sender is unknown to me. [scale from 1 - strongly disagree to 5 - strongly agree]
    - Email, Behaviour: If an email from an unknown sender looks interesting, I click on a link within it. [scale from 1 - strongly disagree to 5 - strongly agree]
    - Email, Behaviour: I don't always click on links in emails just because they come from someone I know. [scale from 1 - strongly disagree to 5 - strongly agree]
- Phishing Self-Reports
  - How would you rate your knowledge on phishing? [scale from 1 - very low to 5 - very high]
  - How would you rate your ability to detect phishing e-mails? [scale from 1 - very low to 5 - very high]
  - How would you rate your level of alertness for phishing attacks? [scale from 1 - very low to 5 - very high]

A.2 Post-Training Questionnaires

- Security Proficiency
  - Theoretical Phishing Test: 15 multiple choice questions.
  - Practical Phishing Test: Classification of 5 emails.
  - HAIS-Q [36]: 5 items from internet use sub scale [3 attitude, 2 behaviour] 6 items from email use sub scale [3 attitude, 3 behaviour] (see pre-training questionnaire for details)
- Demographics & Background
  - What is your age? [in age ranges]
  - What is you highest level of education? [Did not finish high school, high school, associate degree, university degree (Bachelor/Master), PhD or similar]
  - Do you have an IT security related background or is your occupation concerned with IT security? [Yes, studies in Computer Science, IT Security, Cybersecurity; Yes, IT specialist; Yes, other IT security related education or occupation; No, other education or occupation]
  - What is your occupation? [Training/Apprenticeship, University Student, Employee, Civil Service, Self-employed, Unemployed, Retired, Other]
  - BFI-10 Personality Questionnaire [38]: The short 10-item version of the Big Five Inventory is based on the predominant Five-Factor model of personality [34] that includes openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. The mean re-test reliability for the questionnaire is reported with r_{tt}= .56 [38].
- Phishing Self-Reports
  - How would you rate your knowledge on phishing? [scale from 1 to 5]
  - How would you rate your ability to detect phishing e-mails? [scale from 1 to 5]
  - How would you rate your level of alertness for phishing attacks? [scale from 1 to 5]

- SA-13 [14]: The Security Attitude Inventory measures Engagement with Security Measures, Attentiveness to Security Measures, Resistance to Security Measures and Concernedness with Improving Compliance on 13 items. The relaibility measured through the internal consistency varies between Cronbach's α=.69 and Cronbach's α = .81 [14].
- IUIPC-8 [18]: The Internet users' information privacy concerns originally developed by Malhotra et al. [32] and later validated by Gross [18] in an eight-item version measures privacy concerns related to the dimensions control, awareness, and collection. The reliability measured via internal consistency exceeds values of Cronbach's α > .7 [18].
- An open comment field.

A.3 Knowledge Test Questions

This section lists the questions used to ascertain participant's theoretical phishing knowledge.

- What is phishing?
  - A type of malware that infects computers and steals personal data.
  - A type of online advertising that uses deceptive tactics to attract clicks.
  - A type of social engineering used to convince people to buy products they don't need.
  - A type of online scam where criminals send fraudulent messages to trick people into sharing sensitive information.
- What is a risk of falling victim to a phishing attack?
  - Identity theft
  - All these answers are correct
  - Loss of personal data
  - Computer infected by malware
  - Financial loss
- Suppose you have received an email from PayPal asking to reset your password, and the email sender address is one of the following. Which one of these mail addresses looks suspicious?
  - servicepaypal.com
  - servicepaypalgmail.com
  - None
  - paypalpaypal.com
  - supportpaypal.com
- Suppose now you have received an email from Dropbox saying that a user has shared his Dropbox with you. Which one of the following mail addresses looks suspicious?
  - no-replydropbox.com
  - no-replyem-s.dropbox.com
  - None
  - noreply-dropboxnoreplysupport.com
- Among the following sentences, which one would you most likely find in a phishing email?
  - "As part of our routine account maintenance, we kindly request that you confirm your personal information by visiting our website and completing the validation process accordingly."
  - "This is a notification from the Cybercrime Division of your local police department. Our records indicate that your internet connection was used to access illegal content and distribute malware. To avoid prosecution, you must pay a fine of 500 [] within 24 hours. Failure to comply will result in legal action being taken against you."
  - "We noticed some unusual activity on your account and wanted to confirm whether you made the transaction of 500 [] at [Merchant Name] on [Date]. If this was not you, please contact us immediately to resolve the issue."
  - None
- Among the following sentences, which one would you most likely find in a phishing email?
  - "It was a pleasure having you as a customer. Don't hesitate to come back or contact us to the following link if you need our help again."
  - "As a valued customer, we're giving you a special discount! -90% on all our offers, click here to view more!"
  - None
  - "Please visit our official website to update your personal account."
- Among the following sentences, which one would you most likely find in a phishing email?
  - "Thank you for your registration! Click here to see your account details."
  - None
  - "Click here to view the latest collection of our awesome brand!"
  - "Your bank account password has been compromised! If you don't act fast, hackers might steal your money! Click here to reinitialize your password!"
- Among the following sentences, which one would you most likely find in a phishing email?
  - None
  - "Please update your account information to continue using our service, click on the following link"
  - "We are excited to announce our new website and features. Find more by clicking on the following link"
  - "Don't forget to submit your report by Friday, April 9th. Submit with the following link"
  - "Please make sure to verify your account information with the following link"
- Suppose you have received an email that contains one of the following links. Which one looks **legitimate**?
  - https://www.youtube.com/watch?v=dQw4w9WgXcQ&t=3s
  - https://www.you.tube.com/account
  - https://cutt.ly/V8rT54mnJ90
  - None
  - http://google-file-share.c.com
- Suppose attached to a mail you have one of the following files. Which one looks **legitimate**?
  - None
  - paypal_account_details.exe
  - bank_invoice.scr
  - Important_doc.docx
  - bank_account.pdf.exe

- If you fall for a phishing scam, what should you do to limit the damage?
  - Change any compromised password
  - Unplug your computer to get rid of the malware
  - Delete the phishing email
- What should you do as an employee if you suspect a phishing attack?
  - Ignore it
  - Show your coworkers to see what they think
  - Report it so the organization can investigate
  - Open the email and see whether it looks legitimate
- What is a common type of content found in phishing emails?
  - Security alert of suspicious login from an unknown location
  - Advertisements for weight loss supplements
  - Threats of account deactivation or legal action if immediate action is not taken
  - Unsolicited job offers
- The following address is suspicious: googleaccountsupport-gsupport.com. Why?
  - The domain address should contain google.com
  - It should contain no-reply
  - Company names in email addresses always have capital letters, so “google” should be written “Google”
  - It is not suspicious
  - It should be a gmail address

B SCORE CALCULATION AND CATEGORISATION

B.1 Score Calculation Used for Personalisation

The training was personalised based on the phishing proficiency level of the user. To determine this proficiency, scores were calculated based on pre-training questionnaire answers about participants’ background, knowledge, and abilities. We iteratively developed a system of weights internally and later tested and adjusted it in a pre-study with first 10 and then 20 more participants. The goal was to differentiate between knowledge levels (to recognise low proficiency users that profit from education elements) and awareness levels (to identify medium proficiency users that profit best from a training). This process involved putting more weight on fundamental questions (such as knowing what phishing is) and on more direct variables (such as weighing phishing knowledge items higher than HAIS-Q items, which capture more general cybersecurity behaviour).

We calculated a theoretical and practical score, combined together into a total score as follows:

- **Theoretical score:** Attitude and knowledge items
  - 11 HAIS-Q items (as a good predictor for cybersecurity attitudes and intentions [35]), with 0.5 points for highest scale point (5 - strongly agree), 0.3 points for the second highest scale point (4), 0.1 points for the third-highest scale point (3), 0 points for the weakest or second-weakest scale point (2 and 1 - strongly disagree).
  - 3 general questions about the concept of phishing, with 4 points for each correct answer.
  - 9 specific questions about what to do when confronted with potential phishing emails, 2 points for a correct answer.
  - 2 questions relating to reactions (e.g., reporting) to phishing emails, with 2 points for a correct answer, and 1 point for a half-correct answer.
- **Practical score:** Phishing classification task and background
  - 5 exemplary email screenshots that had to be classified into phishing or not phishing, with 2 points for each correct answer.
  - Ability self-report items, with 1 point for highest (5 - very high), 0.5 points for second-highest (4 - rather high), and 0.25 for third-highest scale points (3 - medium) in a 5-point Likert scale for each knowledge, ability, and alertness ratings.
  - Experience with cybersecurity training, with 2 points for participating more than once in training, and one point for participating in training once.
  - Frequency of email client use, with 0 points for "less than once a week" and "approximately once a week", and 1 point for all other answers.
  - Amount of emails received in a typical workday, with 0 points if "Less than 5" was selected, and 1 point for higher amounts.
- **Total score:** Sum of theoretical and practical scores.

B.2 Categorisation

These scores were then used to determine the appropriate level as follows. We employed an initial step to assign participants with low theory score to the low proficiency group, even if they scored well in other areas, as they might lack knowledge in either case and could still benefit from educational elements.

1) If theory score was below 20.5, sorted into *low proficiency*.
2) If total score was below 27.5, sorted into *low proficiency*.
3) If total score was above 39.5, sorted into *high proficiency*.
4) Everyone else sorted into *intermediate proficiency*.

B.3 Calculation Example

Below, in Table 1, we show a calculation for an hypothetical participant that would be assigned to the low-proficiency group to make the calculation process more graspable:

Score Component	Explanation	Score
Theoretical Score		
HAIS-Q items	selected scale point 4 for 6 items = 1.8 points selected scale point 3 for 3 items = 0.3 points selected scale point 1 for 2 items = 0 points	2.1
Questions about the concept of phishing	2 out of 3 questions correct = 8 points	8
Questions about what to do when confronted with potential phishing	4 out of 9 questions correct = 8 points	8
Questions relating to reactions to phishing emails	1 out of 2 questions correct = 2 points	2
Practical Score		
Classification of email screenshots	2 out of 5 correct = 4 points	4
Ability self-report items	selected scale point 4 for 3 items = 0.5 points	1.5
Training experience	Never had training = 0 points	0
Email client use	Daily use = 1 point	1
Amount of emails	<5 emails per day = 0 points	0
Sum Score	**Categorised as low proficiency**	**26.6**

Table 1: Calculation example for a hypothetical participant.

C BACKGROUND TASK AND TRAINING SCREENSHOTS

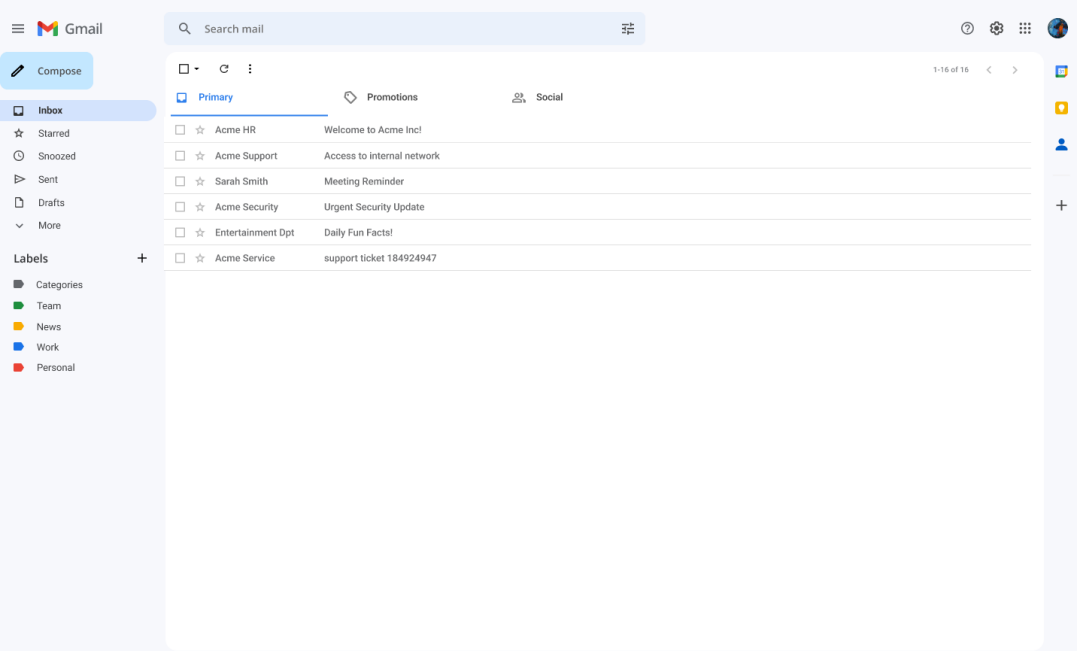


Figure 5: An overview of some of the emails during the background tasks.

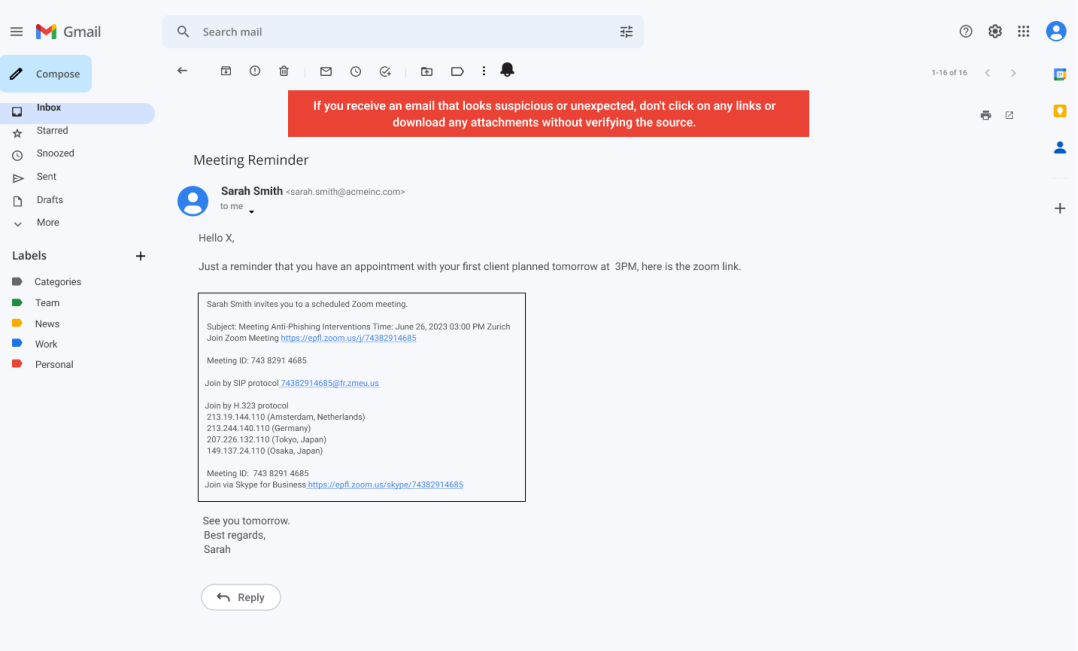
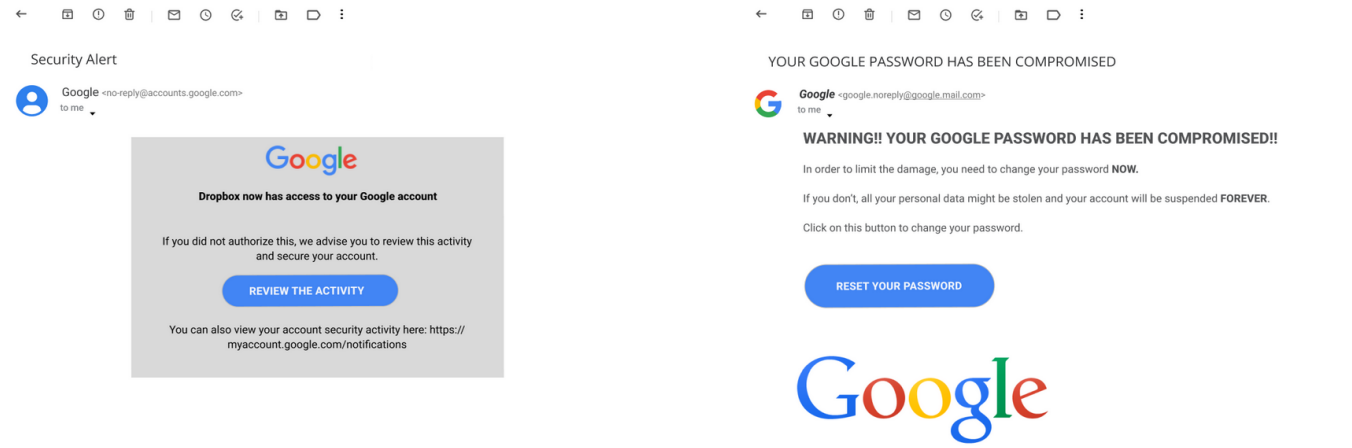


Figure 6: An example of a task with a reminder superimposed on top.



(a) A training element where users are presented with a non-phishing email.

(b) A training element where users are presented with a phishing email.

Figure 7: Training elements

D STATISTICAL RESULTS

D.1 Differences between User Groups

Table 2: Overview of Variables and their Differences Between User Groups. Significant Influences are Bold.

Variable	Test	Result
Email Use Frequency	Fisher's Exact Test	$p = .143$
Daily Emails Received	Fisher's Exact Test	$p = .538$
Daily Spam Received	Fisher's Exact Test	$p = .607$
Professional Mailbox	Fisher's Exact Test	$p = .241$
Security Training	Fisher's Exact Test	$p = .001$
Self-reported Knowledge	Kruskal-Wallis-Test	$H(2, 94) = 9.80, p = .007$
Self-reported Ability	Kruskal-Wallis-Test	$H(2, 94) = 10.50, p = .005$
Self-reported Alertness	Kruskal-Wallis-Test	$H(2, 94) = 8.56, p = .014$
Age	Fisher's Exact Test	$p = .923$
Education	Fisher's Exact Test	$p = .867$
IT Education	Fisher's Exact Test	$p = .001$
HAIS-Q Email Attitudes	Kruskal-Wallis-Test	$H(2, 94) = 8.03, p = .018$
HAIS-Q Internet Attitudes	Kruskal-Wallis-Test	$H(2, 94) = 4.54, p = .104$
HAIS-Q Email Behaviour	Kruskal-Wallis-Test	$H(2, 94) = .51, p = .773$
HAIS-Q Internet Behaviour	Kruskal-Wallis-Test	$H(2, 94) = .47, p = .789$
SA-13 Engagement	Kruskal-Wallis-Test	$H(2, 94) = 4.07, p = .130$
SA-13 Attentiveness	Kruskal-Wallis-Test	$H(2, 94) = 6.94, p = .031$
SA-13 Resistance	Kruskal-Wallis-Test	$H(2, 94) = 7.40, p = .025$
SA-13 Concernedness	Kruskal-Wallis-Test	$H(2, 94) = 1.51, p = .469$
SA-13 Total	Kruskal-Wallis-Test	$H(2, 94) = 9.05, p = .011$
IUIPC-8 Control	Kruskal-Wallis-Test	$H(2, 94) = 6.18, p = .046$
IUIPC-8 Collection	Kruskal-Wallis-Test	$H(2, 94) = 2.27, p = .321$
IUIPC-8 Awareness	Kruskal-Wallis-Test	$H(2, 94) = 6.51, p = .039$
IUIPC-8 Total	Kruskal-Wallis-Test	$H(2, 94) = 7.55, p = .023$

D.2 Personality Traits

Table 3: Coefficients from Linear Regression Model of Personality Traits Affecting Total Pre-Training Score.

Coefficient	Beta	95% CI	p-value
Extraversion	-.43	-1.9, 1.0	$p = .562$
Agreeableness	.45	-1.1, 2.0	$p = .570$
Conscientiousness	.77	-0.85, 2.4	$p = .350$
Neuroticism	.13	-1.2, 1.4	$p = .849$
Openness	.25	-1.2, 1.7	$p = .731$

Table 4: Coefficients from Linear Regression Model of Personality Traits Affecting Total Post-Training Score.

Coefficient	Beta	95% CI	p-value
Extraversion	-1.83	-3.5, -0.13	$p = .035$
Agreeableness	.03	-1.8, 1.8	$p = .978$
Conscientiousness	.24	-1.6, 2.1	$p = .800$
Neuroticism	-1.09	-2.6, 0.42	$p = .155$
Openness	.93	-0.71, 2.6	$p = .261$

D.3 Internal Consistency Metrics

Table 5: Internal Consistency Metrics Indicating the Reliability of Instruments with our Participant Sample

Instrument	Method	Reliability
BFI-10	McDonald's ω^*	.78
HAIS-Q Pre-Training	Cronbach's α	.80
HAIS-Q Post-Training	Cronbach's α	.88
SA-13	Cronbach's α	.80
IUIPC-Control	Cronbach's α	.78
IUIPC-Awareness	Cronbach's α	.75
IUIPC-Collection	Cronbach's α	.88

A value of .70 or higher is in line with other research and acceptable [37].

*We employed McDonald's ω for the BFI-10, as it is a more appropriate method for heterogeneous scales, commonly present for personality traits [19, 33]. The reported measures are still comparable [33].