Understanding Human Aspects in Phishing Detection: The Role of Demographics, Eye Movements and User Experience in Security Software

Fabian Engl, Meret Kristen, Jürgen Mottok
Software Engineering Laboratory for Safe and Secure Systems
OTH Regensburg
Regensburg, Germany
email: {fabian.engl | meret.kristen | juergen.mottok}@oth-regensburg.de

***Abstract*—This paper builds upon a previous study that analyzed phishing detection using eye-tracking data from 103 participants tasked with classifying 18 emails. Additionally, a phishing awareness system (PAS) was introduced, highlighting relevant information for half of the participants. While the original analysis found no significant improvements in detection effectiveness, the eye-tracking data did reveal that participants using the supportive software spent less time examining key phishing indicators. Expanding on these findings, this work incorporates further questionnaire data and a more advanced Area of Interest (AoI) analysis to provide deeper insights. The results indicate that demographic factors such as age, gender, and education have no significant impact on phishing detection. However, industry sectors and weekly screen time did influence performance, particularly in terms of the time required for classification. A qualitative eye movement analysis further revealed distinct AoI hit patterns between participants who correctly classified all emails and those who misclassified more. Additionally, gaze behavior varied based on participants' usability and user experience ratings of the supportive software, highlighting a potential impact for specific user groups, when it comes to phishing detection efficiency.**

Keywords-Phishing; Security Awareness; Eye-Tracking; IT-Security; Usability and UX.

I. Introduction

This paper builds on previous research that investigated phishing detection using eye-tracking analysis [1]. Despite widespread awareness of phishing and its associated risks, these attacks remain a persistent daily threat. The German Federal Office for Information Security (BSI) highlighted in its 2024 IT-Security report that many individuals continue to underestimate the severity of phishing, often realizing the long-term consequences only when it is too late [2]. Phishing attacks typically disguise themselves as legitimate emails or messages to deceive individuals into revealing sensitive information, such as login credentials, financial details, or confidential data. As the volume of emails continues to rise and phishing tactics grow more sophisticated, individuals are becoming increasingly vulnerable. Historically, phishing emails primarily impersonated financial institutions, requesting monetary transfers; however, in recent years, they have shifted towards everyday communications, making these attacks both more pervasive and harder to detect [2].

Given the increasing prevalence and complexity of phishing attacks, equipping individuals with the skills to recognize these threats is more crucial than ever, both in personal and corporate settings. Traditional in-company security awareness trainings - often based on theoretical knowledge - have sparked a debate regarding their effectiveness in preventing such attacks. Some argue that humans are the weakest link in cybersecurity [3] [4] and that dedicated training can significantly improve individuals' ability to recognize phishing threats [5]. However, studies such as the one conducted by Lain et al. suggest that such training has no significant impact on a person's ability to detect phishing emails [6].

Phishing research is typically conducted using questionnaire-based studies [7] [8] [9]. However, questionnaires may not fully capture an individual's decision-making process, often providing an incomplete or distorted picture of the cognitive mechanisms involved in phishing detection. Eye-tracking technology, on the other hand, offers a more precise representation of decision-making processes [10]. This journal paper first conducts a statistical analysis to determine whether and which demographic factors influence phishing email detection and then explores how eye-tracking data can provide deeper insights into decision-making patterns that remain hidden in traditional phishing studies.

Section II reviews recent literature published since the initial study. Sections III to V define and address eight research questions, beginning with statistical tests on questionnaire data and ending with a qualitative analysis of eye movement patterns. These sections also provide a detailed overview of the participants, the technical setup, and the study design. Section VI further investigates the usability and user experience of a software add-on designed to highlight phishing-relevant information. Finally, Sections VII and VIII discuss the study's limitations, summarize key findings on the effectiveness of phishing training, and outline directions for future research.

II. Recent Developments in Literature

The literature review in [1] revealed that previous eye-tracking research studying phishing either relies on relatively small sample sizes or focuses on adaptive mechanisms designed to enhance users' ability to recognize phishing attempts. However, there remained a significant gap in understanding how users engage with available tools and warnings, as well as which phishing indicators they tend to overlook when falling victim to such attacks. To address these gaps, the study in [1] was developed. Since the literature review for that article had to be carried out before the start

of the study and the implementation of the study itself took several months, the literature review did not include articles published after February 2024. In the last 12 months, several new papers have been published studying phishing emails with eye-tracking technology. For this reason, a discussion of these new papers and how their results compare to those found in [1] is included here.

In [11], published in November 2024, the authors examined how individuals visually inspect phishing and legitimate emails. A key hypothesis was that participants would pay more attention to the sender's identification in phishing emails compared to legitimate ones, which was confirmed by the results. However, contrary to expectations, participants were not more likely to misidentify phishing emails; instead, they tended to misclassify legitimate emails more frequently.

The study involved 68 participants, predominantly women (77.9 %), with an average age of 23.91 years (ranging from 18 to 48). More than half of the participants (51.5%) had completed their 12th grade education.

Data was collected using Tobii Pro Fusion Eye-Trackers. A total set of 28 emails was examined, consisting of 13 phishing emails, 13 legitimate emails, and two control emails, each with predefined areas of interest (AOI), including the header of the email, the subject line, the sender's information, the body of the email, the salutation, the links, misspellings, financial indicators, threats, and urgency signals. Each participant was shown 15 randomly selected emails from two groups.

The study analyzed the total number of fixations and the fixation duration (in milliseconds) within each AOI. Statistical comparisons were conducted using Mann-Whitney U-tests. The general setup of the study is very similar to that presented in [1] and the results show that participants spent more time looking at the sender information in phishing emails. Since [1] only tested whether AOI hits on the sender information differ between the group with PAS and the group without, this result will be tested against the original data set from [1] in research question 6, to see whether the collected data is consistent.

Furthermore, the authors of [11] suggest that future research should differentiate between specific phishing characteristics, such as financial content, threats, spelling errors, and urgency cues. This was already addressed in [1]. Another suggestion was to examine the visual inspection patterns of phishing experts and previous victims, which is addressed in research question 7 below.

In [12], a literature review with the search string [phishing AND EEG], [phishing AND "eye-tracking" OR eye-tracking], [phishing AND BCI] in Elsevier ScienceDirect, IEEE Xplore, Research-Gate, Springer, and the ACM Digital Library is presented. Similarly to the literature review in [1], the found papers are compared with regards to participants, types of investigated phishing attacks and results. The examined literature suggests that user personality traits, such as attention control, may have a direct impact on their susceptibility to phishing. The paper describes the controversy surrounding the impact of demographic factors on phishing susceptibility and the limited scope of current studies. It suggests further research to explore other phishing types, assess resilience to multiple attacks, and incorporate advanced AI methods and real-world conditions.

[13] presents an eye-tracking study with $n = 40$ participants and 18 emails. This study explores the effects of visual risk indicators on phishing detection behavior using an eye-tracking experiment, and provides implications for how organizations can effectively integrate and calibrate such indicators to mitigate phishing attempts. It studied how displaying a phishing risk indicator affects visual attention, trust, and time taken to come to a decision. It was discovered that the visual risk indicator has a significant impact on trust, which subsequently influences the behavior of the participants' email responses.

[14] investigates how workload influences an individual's likelihood of falling for phishing attacks, utilizing eye-tracking technology to track how participants read and engage with personalized phishing emails. By combining both quantitative and qualitative approaches, it analyses participants' focus on two key phishing cues: the sender's email address and hyperlink URLs. Results reveal that paying attention to the email sender helps reduce phishing vulnerability, but no link between noticing the actual URL and improved phishing detection was found. In contrast, focusing on the text hiding the links tends to increase phishing risk. These suggestions are addressed in Research Questions 6 and 7.

Lastly, [15] presents an eye-tracking study with 42 participants that focuses on spear phishing. The results show that the participants have shorter total fixation durations on spear phishing emails than on legitimate emails. Phishing training was not shown to have a main effect on eye movement behaviors. Participants tended to focus their attention on the email body, followed by the subject line and sender information, but neglected the sent time.

III. Research Objectives

Several new questions arise from the review of literature published in the last year. Together with further analysis of the data set presented in [1], this gives rise to the following set of research questions:

RQ1 How do demographic differences such as age, gender, and education affect phishing recognition?

RQ2 Are there differences between employees of different industries in regards to effectiveness and efficiency of phishing recognition?

RQ3 Does the ability to recognize phishing emails differ among employees based on their weekly screen time?

RQ4 Does knowing the sender company affect the recognition of phishing emails?

RQ5 Are IT security experts better at detecting phishing emails than laypersons?

RQ6 Do users focus more on the sender when examining phishing emails compared to legitimate emails?

RQ7 How do gaze patterns differ between individuals who correctly identify a high versus a low number of phishing emails?

RQ8 How do gaze patterns differ between individuals who rate the usability and user experience (UX) of the PAS as low compared to those who rate it as high?

Based on these research questions, the following hypothesis were developed:

H1 Age, gender and education level have little to no effect on phishing recognition rates.

H2 Employees of different industries express different levels of phishing recognition efficiency and effectiveness, proportional to their use of email in daily life.

H3 Participants with increased weekly screen-based work hours show higher rates of phishing recognition.

H4 Knowing the sender company will effect phishing recognition rates.

H5 IT security experts are expected to perform better at the phishing recognition task than laypersons.

H6 In line with the results found in [11] it is expected that users focus more on the sender when examining phishing emails.

H7 Individuals who correctly identified more phishing emails used the PAS more compared to the individuals that misclassified phishing emails.

H8 Individuals who rated the usability and especially the UX as high spend more time interacting with the PAS compared to those who rated both low.

IV. Study design

As described in [1], this eye-tracking study was conducted at the University of Applied Sciences in Regensburg (OTH Regensburg) and as part of a service offered by the European Digital Innovation Hub "Digital Innovation Ostbayern" (DInO). DInO offers free consulting services to small and medium-sized enterprises (SMEs) and the public sector (PSEs), especially in Eastern Bavaria. Since IT security training is mandatory for many German companies, this study was designed as an interactive extension to traditional theoretical training.

Beyond corporate use, the study also aimed to help individuals develop a better awareness of phishing emails and improve their ability to detect them. To ensure relevance and familiarity, the phishing emails used in the study were sourced primarily from real interactions. All were genuine phishing attempts, collected from colleagues and relatives. In some cases, minor modifications—such as translations or company name changes—were made to prevent reputational harm to smaller businesses.

Notably, while all participants were exposed to the same phishing emails, half of the group had access to an additional tool called the "Phishing Awareness System" (PAS), which highlighted specific information. This system will be introduced in Section IV-D.

A. Participants

A total of 120 participants took part in the study. However, since the study was also offered as a complementary phishing training, eleven participants opted to participate only in the training without being included in the study. Their recordings were deleted immediately after the session and were not included in the final dataset. Additionally, six participants had to be excluded due to severe visual impairments, as they failed to meet the calibration threshold of $0.75°$, primarily due to extreme diopter levels or incompatible glasses and contact lenses. Before beginning, all participants filled out a consent and demographic form.

In the final dataset, 103 participants remained, of whom 36.89% were female ($n = 38$) and 63.11% were male ($n = 65$), with an average age of 32.81 years. Among them, 52 had access to the Phishing Awareness System (PAS), while 51 relied solely on the email content for their decisions. 91.26% ($n = 94$) reported knowing what phishing emails look like, and 60.19% ($n = 62$) had attended at least one phishing training session in the past. Additionally, 38.83% ($n = 40$) received phishing emails daily, 28.85% ($n = 30$) weekly, 8.65% ($n = 9$) monthly, and 23.30% ($n = 24$) rarely or never.

A closer look at participants' educational backgrounds revealed an atypical distribution. Based on the German education system, four educational attainment groups were identified:

- 57 participants had a general or subject-specific university entrance qualification (German: Abitur/Allgemeine oder fachgebundene Hochschulreife).
- 10 participants had a technical college entrance qualification (German: Fachhochschulreife).
- 23 participants had a general secondary education diploma (German: Realschulabschluss/Mittlere Reife).
- 11 participants completed basic secondary schooling (German: Hauptschulabschluss).
- 2 participants reported other forms of schooling.

Participants were also asked about their professional qualifications and degrees. Since this was a multiple-choice question, the number of responses exceeds the total number of participants:

- 52 participants had completed an apprenticeship or professional training (German: Berufsausbildung)
- 31 participants had a bachelor's degree
- 32 participants had a master's degree
- 2 participants had a PhD

This distribution is particularly noteworthy since eye-tracking studies are often academically biased, predominantly consisting of students and university employees as well as teachers [16] [17] [18]. The fact that over half of the participants had completed an apprenticeship or professional training highlights not only the scale but also the diversity of this study. A further demographic analysis showed that 85.44% ($n = 88$) of participants were employed, while 14.56% ($n = 15$) were self-employed. Among all, 66.99% ($n = 69$) worked full-time, 18.45% ($n = 19$) worked part-time, and the remaining participants reported other forms of employment, including

apprenticeships or mini-jobs. The average weekly working hours were 35.90 ($min = 8$, $max = 55$, $std = 9.05$), with participants spending an average of 27.16 hours in front of a computer screen ($min = 0$, $max = 55$, $std = 12.21$). The average work experience was 14.27 years ($min = 0$, $max = 45$, $std = 12.99$).

B. Technical Setup

Up to nine Tobii Pro Fusion eye-trackers were used to record the data, with a recording frequency of $250Hz$. Participants were positioned approximately 65 cm from a 21-inch monitor set to a resolution of 1920×1080 pixels, running at $60Hz$. These specifications align with the quality analysis and recommendations in [19]. Following these guidelines, participants were instructed to remain still during the recording and avoid head movements.

The study was conducted using Tobii Pro Lab software (Version 1.232.52758) and employed the Tobii I-VT fixation filter. The Tobii Pro Fusion devices operated on firmware version 1.19.22.

C. Stimuli

To enhance the study design, emails were categorized into three groups, each representing a common type of phishing attack. A total of 18 emails were included in the study, evenly distributed as follows:

- **Control emails**
  Legitimate, harmless emails, such as notifications from energy providers or PayPal.
- **"Badly made" phishing emails**
  Contained multiple red flags, such as cryptic sender addresses or severe misspellings, making them easier to identify.
- **"Well-crafted" phishing emails**
  More sophisticated attempts with only minor misspellings, subtle anomalies, or unusual attachments, for example Word documents containing macros.

These distinguishing features, which allow for the classification of phishing emails, will be referred to as phishing markers throughout the study. Each email category was further divided into three common phishing attack techniques:

- Two emails with attachments containing relevant documents, primarily invoices or monthly billing statements.
- Two emails urging the recipient to click a link to complete an action, such as reactivating an account.
- Two emails requesting money, either through a direct demand or an implicit threat of financial consequences.

To create a realistic testing environment, emails were displayed within a typical Outlook email interface. Outlook was chosen because it is among the most widely used email clients [20] and often used in corporate settings.

D. Phishing Awareness System

As mentioned earlier, this eye-tracking study followed an in-between-subject design, with one group having access to a prototype of the Phishing Awareness System (PAS). This system aggregated and displayed key information to assist in identifying phishing attempts. It highlighted critical elements such as the sender domain, URLs within links, and attachment types, helping to expose spelling errors and other suspicious indicators. An illustration comparing the same email with and without the PAS system is shown in Figures 1 and 2.

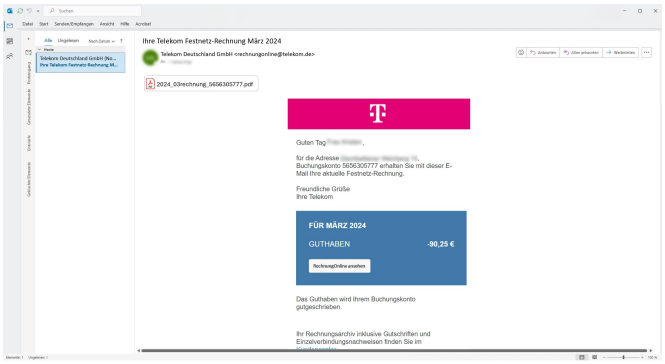

Figure 1. Email from the control group containing an attachment without the PAS.

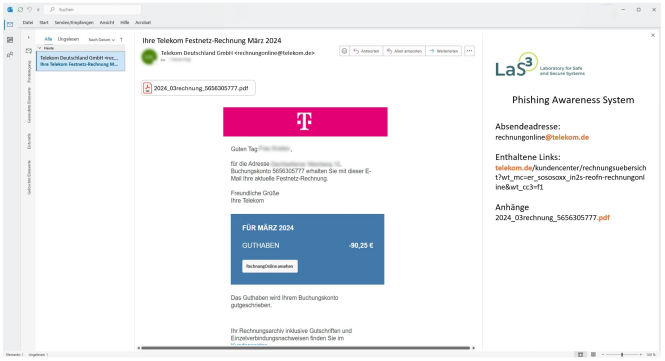

Figure 2. Email from the control group containing an attachment with the PAS at the right side of the screen.

Participants in the PAS group were introduced to the tool and its functionality during the study briefing. However, in order to prevent potential bias, they were not required to use it.

A secondary objective of this study was to evaluate whether the PAS improved participants' accuracy and efficiency in detecting phishing emails, as well as assessing their perception of its usefulness. Section VI will provide a detailed analysis of usability and user experience related to the PAS.

E. Areas of Interest

To analyse participant gaze patterns more effectively, Areas of Interest (AoIs) were predefined. These AoIs represent specific screen regions crucial for determining whether an email is phishing or legitimate. They were drawn over key phishing markers in each email, allowing for the aggregation of eye movements within these targeted areas [21]. By using predefined AoIs, the study systematically examined where

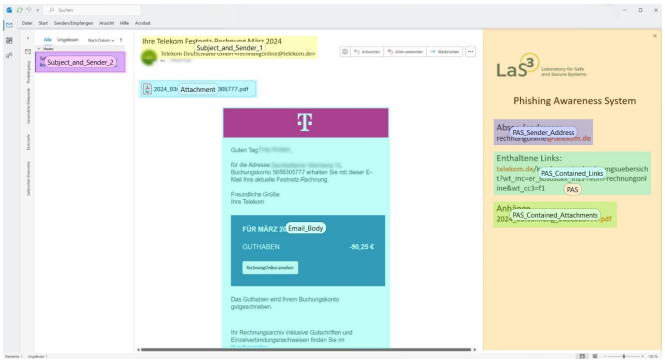

Figure 3. Highlighted AoIs for the email from the control group containing an attachment with the PAS at the right side of the screen.

participants focused their attention and how gaze behavior differed between groups.

Figure 3 provides an overview of these AoIs. In this study, four distinct types of AoIs were defined:

- **Sender Address and Email Subject**
  This information appears twice within the Outlook environment — once at the top of the email and again in the preview pane on the left-hand side. It includes the senders email address, its domain, and the email subject.
- **Email Body**
  This AoI covers the main content of the email, including all text, embedded links, and any other relevant details.
- **Attachment**
  Attachments are typically displayed between the sender information and the email body. This AoI captures the attachment name, file extension, and its icon, all of which provide visual cues about the file type.
- **PAS**
  This AoI is exclusive to the PAS group and consists of: One large area covering the entire PAS interface and three smaller AoIs highlighting the sender domain, included URLs, and attachment details within it.

F. Study Environment and Methodological Challenges

Since this study was also integrated into existing IT security training programs for SMEs and PSEs, it required a different approach compared to traditional eye-tracking studies conducted in laboratory settings. The challenges between these two environments differ significantly, with mobile studies being inherently more complex, particularly when participants have no prior experience with eye-tracking technology [22].

One of the primary concerns in mobile eye-tracking studies is data quality, which is influenced by two key factors: External distractions where Participants may be interrupted by background noise, other participants, or changes in the study environment and technical as well as environmental factors such as poor lighting conditions, calibration problems, and recording errors [19], [22], [23].

To ensure reliable data collection, the eye-tracking system was calibrated to each participant before the study began. Due

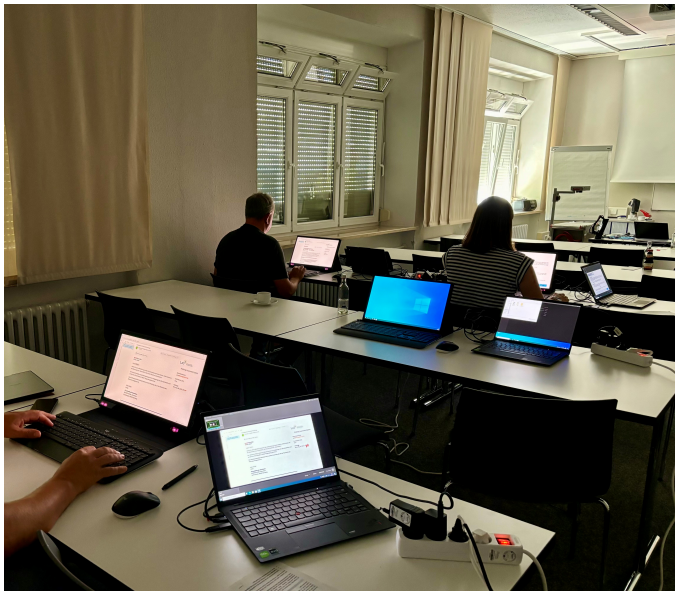

Figure 4. Exemplary study setup for conducting eye-tracking studies in a workshop format.

to the study's relatively short duration (average of $6:40$ minutes), re-calibrations were not performed between stimuli. However, a strict calibration and validation threshold of $0.75°$ was set, and any participant failing to meet this standard was excluded from the study.

To minimize distractions and external influences, several measures were implemented: Firstly, the laptop screen was turned away from participants to prevent distractions. Furthermore, participants were seated directly behind each other to obstruct the view of other screens. Secondly, direct and overhead lighting was turned off and blinds were closed whenever possible to reduce glare. Figure 4 illustrates the typical setup used during workshops.

Beyond technical and environmental factors, participant behavior also played a significant role in data quality. Despite clear instructions to ask questions only during the introduction, some participants raised concerns mid-study, often triggering a chain reaction where others looked away from their screens to listen. In rare instances, discussions emerged among participants, particularly when encountering unusual or suspicious emails. When this occurred, the conductors intervened as discreetly and quickly as possible to minimize disruptions.

For future studies, introducing dedicated breaks between stimuli for questions and short rest periods could be beneficial and combat such behavior. This would allow participants to clarify doubts without disrupting the study flow and help prevent eye strain—an issue raised by participants who needed more time to process all emails.

Despite these challenges, the study demonstrates that parallelism-by-design can enable efficient eye-tracking studies in workshop settings with multiple participants at a time. This was achieved by relying on questionnaires for triangulation, allowing study conductors to oversee multiple sessions si-

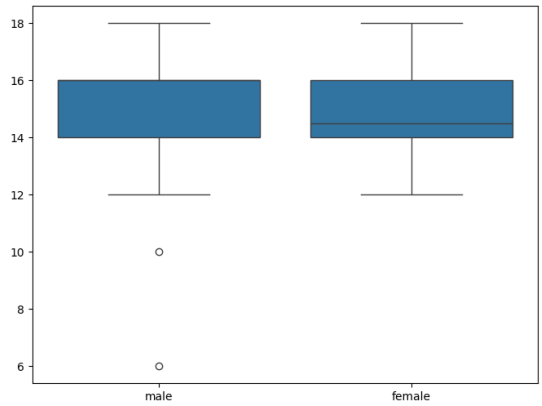

Figure 5. Total number of correctly identified emails for male and female participants.

multaneously. This would not be possible when using think-aloud-protocols or requiring input and validation through the researchers during the study. Self-paced digital instructions further helped the participants progress at their own speed and take out additional stress - which could potentially even introduce bias. All introductions were integrated into the Tobii Pro Lab project, ensuring that everybody received the same information. This had the additional benefit that gaze recordings could be reviewed post-study to verify whether participants actually read the provided instructions.

By implementing these strategies, the study balanced data collection challenges with the practical constraints of real-world IT security training environments, enabling researchers to monitor a higher number of participants while maintaining data integrity.

V. Results

Since no significant differences in phishing detection were found between the group with the Phishing Awareness System (PAS) and the group without it [1], this section further analyses possible correlations by testing demographic differences across the entire dataset, without differentiating between participants with or without PAS.

To test RQ1, a Shapiro-Wilk test [24] revealed that the dependent variable "correctly identified emails" was not normally distributed within the "male" and "female" groups. Therefore, a Mann-Whitney U-test was used, which found no significant differences in the number of correctly identified emails between the two groups at $\alpha = 0.05$ ($z = 1473.00$, $p = 0.099$, $r = 0.16$). Figure 5 displays the results for male and female participants, respectively. Furthermore, no significant group differences were found when comparing the total time taken to complete the task.

Secondly, the results were compared based on participants' highest level of general education. The assumption of normality was assessed using Shapiro-Wilk tests, which revealed non-normal distributions across all groups. Given the lack of normality, non-parametric statistical tests were used. The Kruskal-Wallis test [25] was applied to assess overall group differences, with Mann-Whitney U tests used for post-hoc comparisons. Due to tied ranks in the dataset, p-values were approximated, and continuity correction was applied. Additionally, a Bonferroni correction [26] was used to adjust p-values for multiple comparisons. The Kruskal-Wallis test showed no statistically significant differences between educational attainment groups with regard to Correctly Identified Emails Total ($\chi^2(3) = 3.72$, $p = 0.293$, $\eta^2 = 0.01$). This suggests that educational attainment had a negligible effect on email identification accuracy. Post-hoc Mann-Whitney U tests with Bonferroni correction confirmed the absence of significant differences between any pair of groups, as shown in Table I.

TABLE I. Pairwise Mann-Whitney U-test results for educational levels

	Abitur	**Fachhoch-schulreife**	**Realschul-abschluss**	**Hauptschul-abschluss**
Abitur		$p = 1.000$	$p = 1.000$	$p = .604$
Fachhoch-schulreife	$p = 1.000$		$p = 1.000$	$p = 1.000$
Realschul-abschluss	$p = 1.000$	$p = 1.000$		$p = 1.000$
Hauptschul-abschluss	$p = .604$	$p = 1.000$	$p = 1.000$	

However, statistically significant differences were observed when analysing the time participants needed to complete the study. The Kruskal-Wallis test revealed values of $\chi^2(3) = 15.10$, $p = 0.002$, $\eta^2 = 0.12$, indicating a significant difference at $\alpha = 0.05$ with a moderate effect size.

Pairwise comparisons using Mann-Whitney U tests showed significant differences at $\alpha = 0.05$ between the 'Abitur' and 'Hauptschulabschluss' groups ($p = 0.015$), as well as between the 'Fachhochschulreife' and 'Hauptschulabschluss' groups ($p = 0.026$). These comparisons were conducted with approximated p-values, continuity correction, and a Bonferroni correction applied by multiplying the p-value by the number of tests performed. The results are shown in Figure 6.

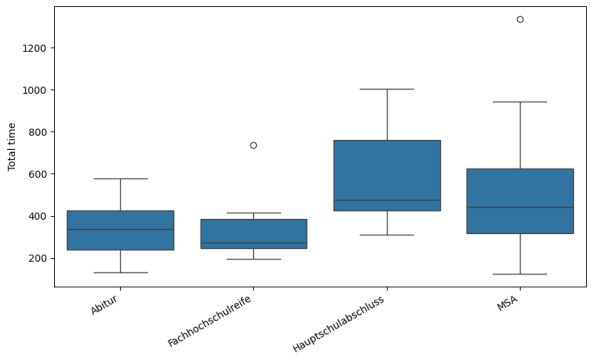

Figure 6. Total time needed to complete the study by highest general education degree

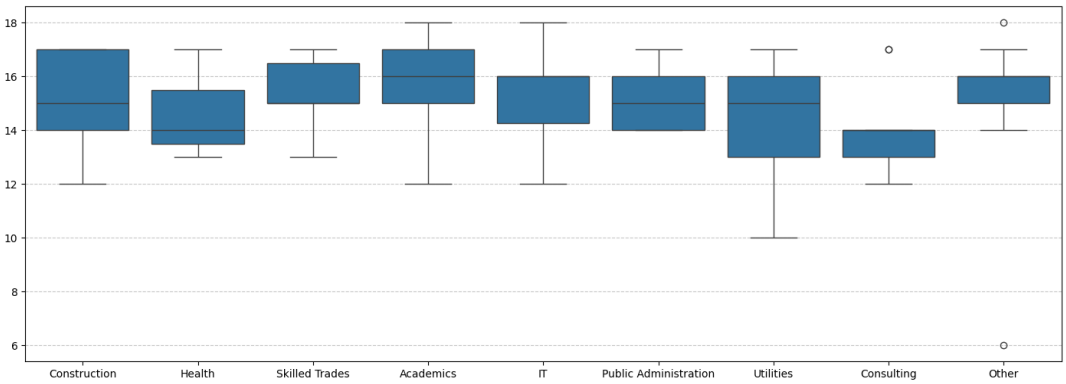

Figure 7. Total number of correctly identified emails by industry sector.

Lastly, the relationship between the number of correctly identified emails and age was examined. The data is described using the median (Mdn), interquartile range (IQR), and sample size (n): For the number of correctly identified emails, $Mdn = 15.00$, $IQR = 2.00$, and $n = 103$. For age, $Mdn = 32.00$, $IQR = 18.50$, and $n = 103$. Both variables are not normally distributed as accessed with Shapiro-Wilk tests: For the number of correctly identified emails the test yields $W = 0.90$, $p < .001$, which is significant at $\alpha = 0.05$, indicating that the sample is not normally distributed. For age, the test yields $W = 0.92$, $p < .001$, which is significant at $\alpha = 0.05$ and again indicates that the sample is not normally distributed. Due to the non-normal distribution of both variables, median (Mdn) and interquartile range (IQR) were used to describe the sample. As a result, a non-parametric test was conducted. Of the two popular non-parametric correlation analyses, Spearman's [27] and Kendall's [28], the latter is considered more conservative (i.e., more likely to not identify significance when it does not exist). Therefore, Kendall's correlation was used in this analysis. Given the presence of ties in the data (i.e., multiple measurements of one variable with the same value), the p-value was approximated. The Kendall's correlation test showed no significant correlation between age and the number of correctly identified emails at $\alpha = 0.05$, with $z = -0.14$, $p = 0.889$, and $r = 0.01$. These findings suggest that there are no statistically significant relationships between email identification accuracy and participants' age, gender, or highest level of general education. Despite some variation in median scores, effect sizes were negligible, and no pairwise comparisons reached statistical significance. The amount of time required to complete the study varied significantly based on participants' highest general education level, but age and gender did not have an effect. These results imply that education plays a significant role in phishing detection, whereas demographic factors such as gender and age do not.

For the second research question, participants were asked to state the industry sector they work in. Any sector that was listed less than three times is listed under "Other", to allow for more accuracy in the statistical tests. To test whether the groups differ in effectiveness and efficiency of phishing detection, it was tested whether the dependent variables "number of correctly identified emails" and "total time spent for the task" were normally distributed. Since this was only the case for 7 out of the 9 groups, non-parametric tests were employed. Here, Kruskal-Wallis test should be used to test for group differences, while Mann-Whitney-U tests with adequate Bonferroni correction may be used as post-hoc tests. For the latter, as there are sample sizes of each two groups are higher than 20, the p-value can be extracted very well from an approximation. Due to unequal sample sizes for both groups, continuity correction is applied. For post-hoc tests in general, the p-values must be adjusted since multiple tests are calculated on the same data. Here, Bonferroni correction is used, which means that p-values are multiplied by the number of pairwise comparisons tests. The Kruskal-Wallis test showed no significant difference in effectiveness at $\alpha = 0.05$ with merely no effect, as shown by $\chi^2(8) = 8.05$, $p = .428$, $\eta^2 = .00$. This can also be seen in Figure 7. However, for the efficiency, a significant difference between the groups was detected. Further pairwise testing with Mann-Whitney U-tests confirmed that the sectors "IT" and and "Construction" differed significantly, shown by a p-value of .015, as well as a significant difference between the groups 'IT' and 'Utilities', shown by a p-value of $p = .003$ after Bonferroni-Correction. None of the other groups showed significant differences in efficiency. The differences between the three relevant industry sectors are shown in Figure 8.

This shows that employees of companies in the IT sector need significantly less time to decide whether an email is legitimate or not than employees in the construction or utilities sector. This effect might be due to familiarity with emails and phishing attempts, advanced knowledge on how possible email scams can look like, and overall confidence in working with a computer. While the industries did not differ in effectiveness of phishing recognition, a difference in efficiency is a good starting point and it should be further analyzed if and how

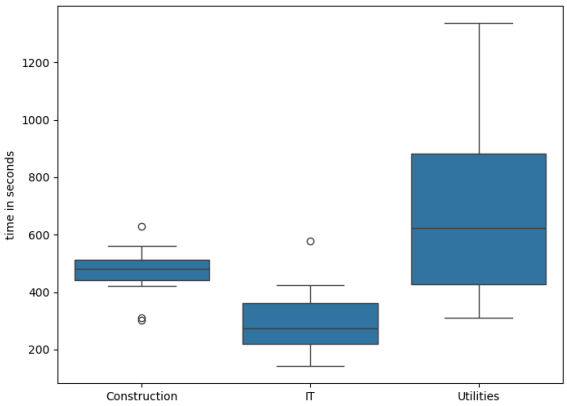

Figure 8. Total time spent on the email sorting task for industry sectors

employees of other industries could be enabled to catch up to the level of expertise shown by employees in the IT sector.

To further evaluate this difference, research question 3 tested whether the weekly screen-based work hours had an effect on the effectiveness and efficiency of phishing recognition. Since employees in the IT-sector naturally spent more of their weekly work hours in front of a computer screen than employees in the Utilities or Construction sector, this is to be expected. And, similarly to RQ2, it was found that while weekly screen-based work hours have only a negligible effect on the total number of correctly identified emails, it has a significant effect on the amount of time needed to complete the task. Since none of the variables are normally distributed, Kendall's correlation analysis was employed and showed that both the total weekly screen-based work hours and the relative weekly screen-based work hours (in relation to total weekly work hours) are significantly correlated to the total amount of time needed to complete the task, as shown by values of $z = -3.13$, $p = .002$, $r = .22$ and $z = -2.71$, $p = .007$, $r = .19$ respectively. This is shown in Figure 10. Furthermore, work experience measured in years had no effect on the efficiency and effectiveness of phishing recognition. To double-check, it was tested whether the number of correctly identified emails and the time spent on the task were correlated, but this was not the case.

Research question 4 answers whether previous knowledge of the sender affects the recognition rates of phishing emails. To test this, the question "From which of the following companies have you already received emails (newsletters, etc.)?" was implemented into the questionnaire for each company presented in the stimuli. Afterwards, the data "correctly classified or not" and "previously known sender or not" were compared for each participant and each email stimulus. A chi-square test of independence (also called a chi-square contingency test) [29] was used to check whether the two binary variables are statistically related. The test revealed a Chi-square statistic of 2.96 and a p-value of 0.085, thus no significant association could be found. Figure 9 shows that participants recognized phishing emails from known senders slightly better than those from unknown senders, but not enough to reach statistical significance.

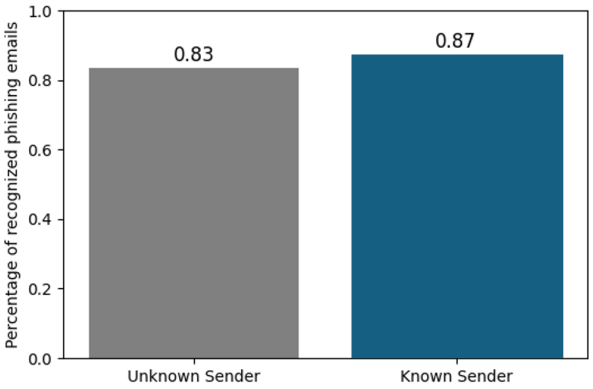

Figure 9. Proportion of correctly identified phishing emails depending on whether the sender was previously known to the participant

To answer research question 5, participants were asked to indicate their agreement or frequency of behavior based on the statements shown in Section II. Responses were recorded using binary values or a 5-point Likert scale whenever suitable, with the values Never, Rarely, Sometimes, Often, Always. The 5-point scale allows for a nuanced assessment of participant behavior rather than a binary yes/no response. Statements are formulated in the first-person to enhance self-reflection and reduce response bias.

TABLE II. QUESTIONNAIRE TO ASSESS PARTICIPANTS IT-SECURITY KNOWLEDGE LEVEL

Statement	Response
I am familiar with the appearance of phishing emails and can identify examples of suspicious characteristics.	yes/no
I use the same password for multiple accounts.	yes/no
I use multi-factor authentication whenever possible.	yes/no
When an update for software or operating systems is available, I install it immediately.	Likert scale (1–5)
I verify the sender's email address before clicking on a link in an email.	Likert scale (1–5)
I check the URL before clicking on a link in an email.	Likert scale (1–5)
I verify the format of attachments before opening them.	Likert scale (1–5)
I open attachments from senders I do not know.	Likert scale (1–5)

For evaluation, the answer "yes" was translated to the numerical value "1" and the answer "no" to "0", except for the question "I use the same password for multiple accounts.", where the value 1 was given to the answer "no" and the value 0 to the answer "yes". This way, a higher score represents a deeper understanding and internalization of IT-security awareness actions. Similarly, for the questions with a Likert scale response, the values were translated as $0 =$ Never, $0.25 =$ Rarely, $0.5 =$ Sometimes, $0.75 =$ Often, $1 =$ Always, except for the last question were the values are reversed in order for the higher score to represent a higher level

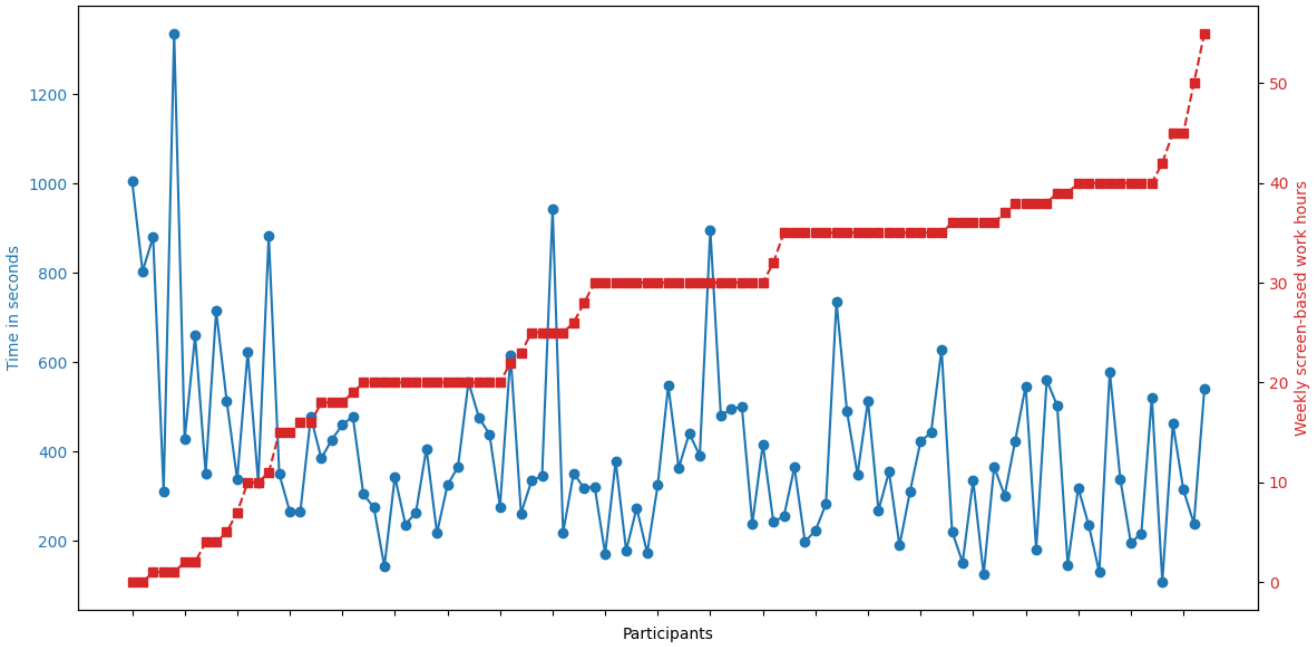

Figure 10. Total time spent on the email sorting task versus weekly screen-based work hours

of IT security awareness. Using these numerical values, the maximum attainable IT-security awareness score was 8, and the minimal score was 0. The mean score was 6.4 ($min = 2.5$, $max = 8$, $std = 1.26$), with a median score of 6.75. The IT-security knowledge level of participants is displayed in Figure 11. The scores varied slightly between industries sectors, with employees in the IT sector showing slightly higher scores than employees in the construction or utilities sector, but not enough to reach statistical significance. Similarly to before, Kendall's correlation test was not able to detect a correlation between the level of IT security knowledge and the number of correctly identified emails. Only a correlation between the IT security knowledge level and the time needed to complete the task was detected ($z = -2.94$, $p = .003$, $r = .20$). Grouping the participants into IT security experts (25th percentile) and novices (75th percentile) shows no differences in number of correctly identified emails (see Figure 12).

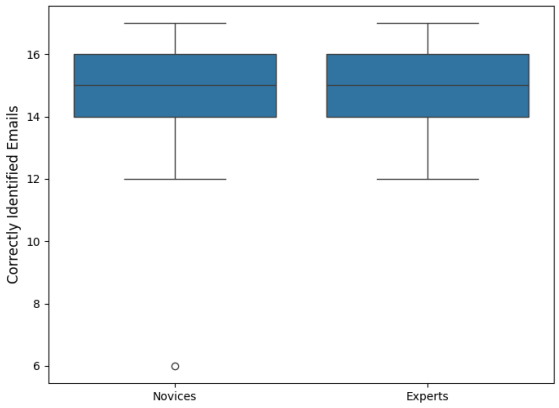

Figure 12. Correctly identified emails by IT security knowledge level

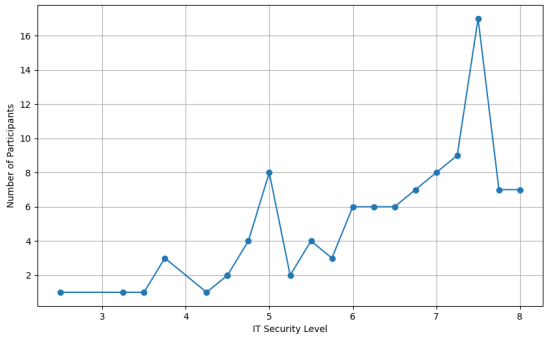

Figure 11. IT security awareness knowledge level of participants

To answer RQ6: it was found that the AOI hits on the Subject and Sender differ significantly between phishing emails and legitimate emails. However, the study was not able to replicate the results found in [11]. On the contrary, it was found that users focused more on the sender when examining legitimate emails than when examining phishing emails. The total AOI hits on all Subject and Sender AOIs were combined (including the PAS-Sender-Address, where the sender address was displayed in the PAS) and it was tested whether these AOI hits differ between phishing emails and legitimate emails. A significant difference was detected by a Mann-Whitney U-test with the values $z = 454328.00$, $p < .001$, $r = .15$, showcasing a significant difference at $\alpha = 0.05$ with small effect. The median for AOI hits on legitimate emails was $Mdn = 494.00$, as compared to a median of $Mdn = 314.00$ for the group of phishing emails. This is shown in Figure 13. This effect might

be explained by the difference in data sets between the two studies. It might have been the case that the phishing emails were easy to spot for the participants, whereas the legitimate ones proved to be more of a challenge. Participants expectancy to be "fooled" could have played a role in their skepticism towards legitimate emails. To test this, it was tested whether participants tended to misclassify legitimate emails more often than phishing emails. In [11], the authors found participants to be more likely to misclassify legitimate emails. The same is the case here, where a total of 85% of phishing emails were recognized correctly, in contrast to only 80% of legitimate emails being recognized as such.

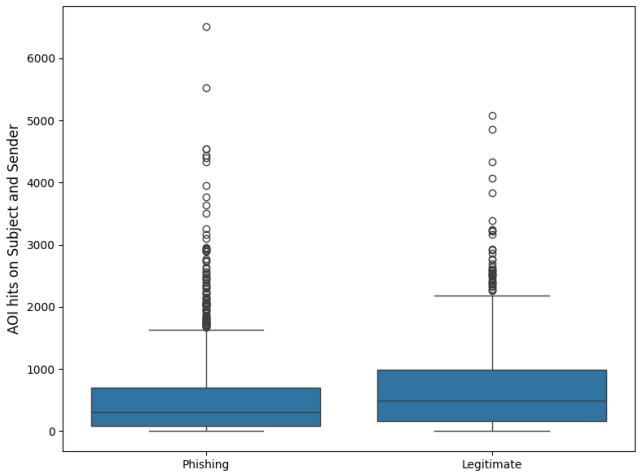

Figure 13. AOI hits on the Subject and Sender Area for phishing emails and legitimate emails

RQ7 builds upon research questions four and five from the original paper [1], offering a deeper analysis of eye-tracking-specific metrics with a focus on the presence of the PAS and its influence on the time spent examining phishing markers. The original paper's AoI analysis indicated that participants with access to the PAS could accurately identify phishing emails equally as efficient while spending less time examining the relevant areas compared to those without the add-on. However, this evaluation was conducted at the group level, without analysing individual participants or emails. Therefore, RQ7 seeks to explore how individual gaze patterns differ between participants who correctly identified most phishing emails and those who misclassified more.

To achieve this, a qualitative analysis is conducted using scarf plots. These visualizations - which are becoming increasingly popular in eye-tracking studies - allow for aggregating gaze movements over time, particularly between AoIs [30].

To compare data at the participant level, appropriate groups must first be defined. Since RQ7 focuses on extremes — participants who correctly identified all phishing emails and those who struggled the most — the groups are determined using quartiles. Examination of the 5th and 95th percentiles for correctly identified phishing emails shows $Q_{0.05} = 8$ $(n = 10)$ and $Q_{0.95} = 12$ $(n = 22)$.

Figure 14. Scarf Plot for the off-brand shoe store email: Visualizing AoI Transitions between phishing markers from participant NOT using the PAS. Participants 1 to 3 are within the $Q_{0.05}$ and participants 4 to 16 are within the $Q_{0.95}$ of correctly identified emails.

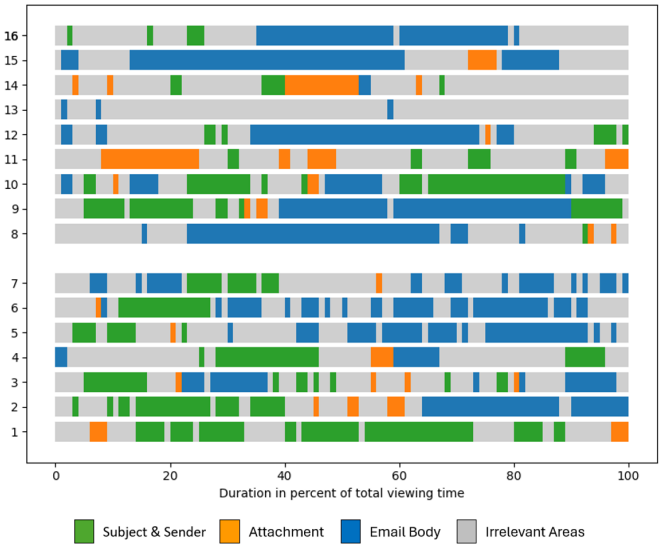

Figure 15. Scarf Plot for the off-brand shoe store email: Visualizing AoI Transitions between phishing markers from participant using the PAS. Participants 1 to 7 are within the $Q_{0.05}$ and participants 8 to 16 are within the $Q_{0.95}$ of correctly identified emails. AoI hits from the PAS and the email itself are combined.

In this case, the 95th percentile consists entirely of participants who correctly identified all 12 phishing emails. In contrast, the 5th percentile group misclassified at least one-third of the phishing emails. Table III provides an overview of participants within the $Q_{0.05}$ range and the phishing emails they misclassified. For clarity and readability, the original participant IDs have been omitted, and participants are renumbered sequentially starting from 1. For all following scarf plots the two groups - with and without the PAS - are separated by

TABLE III. PARTICIPANTS WITHIN THE $Q_{0.05}$ OF FALSELY IDENTIFIED EMAILS. THE PHISHING EMAILS THEY FELL FOR ARE MARKED AS "X"

	Harmful Attachment				Harmful Link				Injunction to send money			
Study Design	Shoe store	Zalando	Mediamarkt	Pustet	Edeka	GMX	DB	iCloud	Schufa	Amazon	Spotify	DHL
Without PAS	X			X			X				X	
Without PAS	X		X	X			X					
Without PAS			X	X						X		X
With PAS	X				X		X				X	
With PAS			X	X			X	X				
With PAS	X						X	X		X		
With PAS		X			X		X	X	X	X	X	X
With PAS	X			X	X		X	X				X
With PAS	X	X	X				X					
With PAS	X		X		X			X				
Sum	**7**	2	5	5	4	0	**8**	5	1	3	3	3

a blank row, with the group that misclassified the most emails listed at the bottom.

All original data, including participants' responses, the stimuli used, and the raw eye-tracking data, can be found on Zenodo (see Section VIII).

Upon reviewing the table, two emails stand out: one from a no-name shoe store and another from the german railway operating company Deutsche Bahn (DB). These were mistakenly classified as legitimate by 7 and 8 out of the 10 participants, making them the focus of the qualitative analysis.

The first of the two phishing emails contained a .zip attachment and a misspelled email address, making it a "bad" phishing email according to the study design.

When analysing the group without the PAS, a clear trend emerges: participants who correctly identified all phishing emails spent more time examining the email body, often scanning this AoI for large sections at a time (see Figure 14). However, participants 5, 15 and 16 stand out in particular, as they spent considerable time looking at irrelevant areas of the screen, areas that could not have contributed to their decision-making. Among those who misclassified the email as legitimate participant 3 stands out. He spent most of his time focusing on the .zip attachment, suggesting that he recognized the potentially harmful file type but did not consider it sufficient enough evidence of a phishing attempt. Interestingly, Participants 1 and 2 ignored the attachment entirely, with Participant 2 not even looking at relevant areas at all.

However, it has to be noted that during the training, several employees emphasized that sending files as a .zip archive is still common practice in small and medium-sized enterprises. Many participants mentioned that in their daily work, they would have reached out or asked a colleague for clarification before making a judgment. Since this option was unavailable in the study, most leaned toward classifying the email as legitimate rather than fraudulent.

A different pattern emerged in the group with the PAS. Here, participants had access to both the email content and additional information from the PAS, highlighting phishing markers. For visualization reasons the scarf plots combine AoIs hits from both the email and the PAS, meaning that participants could examine attachment details within the email or through the PAS, with both being represented as one in the diagram. Participants who correctly identified all phishing emails spent longer periods examining AoIs, switching mainly between different types of information. In contrast, participants who misclassified more emails exhibited fragmented AoI patterns, with frequent short glances at phishing markers (see Figure 15). This suggests they may have mistrusted the PAS and cross-referenced the highlighted phishing markers with the original email content to verify the information manually.

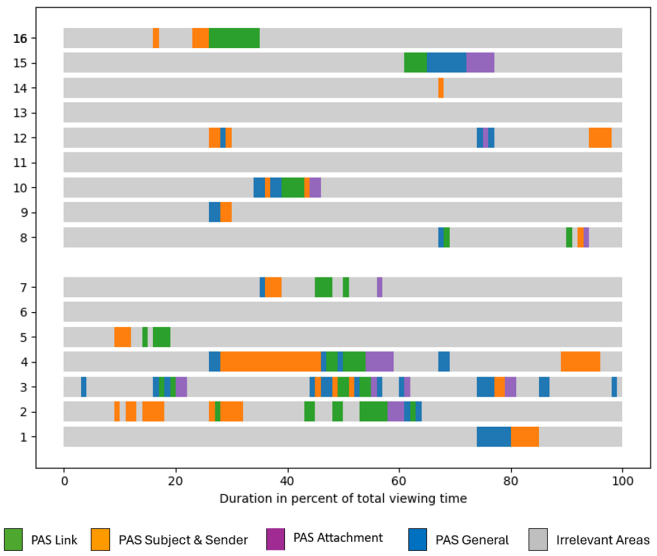

Figure 16. Scarf Plot for the off-brand shoe store email: Visualizing AoI Transitions within the PAS. Participants 1 to 7 are within the $Q_{0.05}$ and participants 8 to 16 are within the $Q_{0.95}$ of correctly identified emails.

However, this trend is not universal. Figure 16 visualizes AoI hits specifically within the PAS. Participants 2 to 4 engaged in the verification process by spending only short periods reviewing the phishing markers highlighted by the PAS, while others barely interacted with the PAS at all. This indicates that participants who misclassified the email either did not trust the PAS or preferred to verify the details manually, if they used the add-on at all. On the other hand,

participants 8 to 16, despite rarely using the PAS, all except one looked at the attachment information at least once. Even this brief engagement with this information may have been enough to help them recognize the email as a phishing attempt.

The second phishing email, from Deutsche Bahn, claimed that the recipient's account would be deactivated unless they took action and clicked on a re-activation link. Unlike the previous email, this one included the company logo and had only a minor misspelling in the sender domain (missing the letter "e": support@**deutsch**-bahn.de). Due to its more convincing appearance, it was categorized as a "good" phishing email in the study design.

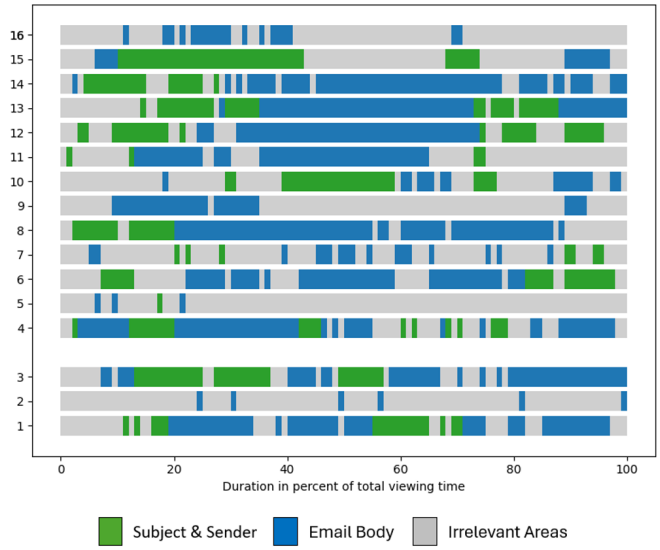

Figure 17. Scarf Plot for the Deutsche Bahn email: Visualizing AoI Transitions between phishing markers from participant NOT using the PAS. Participants 1 to 3 are within the $Q_{0.05}$ and participants 4 to 16 are within the $Q_{0.95}$ of correctly identified emails.

For participants without the PAS, Figure 17 shows no clear difference between those who fell for the email and those who correctly identified it as phishing. As with the previous email, Participant 2 barely looked at any relevant areas, which might suggest he did not take the study seriously or maybe was overwhelmed with the task. The same is true for participant 16 in this scarf plot.

However, when analysing the group with the PAS, a similar trend to the previous phishing email emerges. Participants who misclassified the email as legitimate exhibited more frequent, short, and abrupt switches between different AoIs (see Figure 18).

Interestingly, when focusing solely on PAS usage, participants who correctly identified the email as phishing showed significantly higher engagement with the PAS compared to the previous "bad" phishing email (see Figure 19). This suggests that the PAS is particularly helpful in more subtle cases where crucial phishing markers are easy to overlook. Additionally, the increase in PAS usage toward the end of the decision-making process indicates that participants trusted the information provided by the PAS, using it either as the basis for their decision or at least as a final verification.

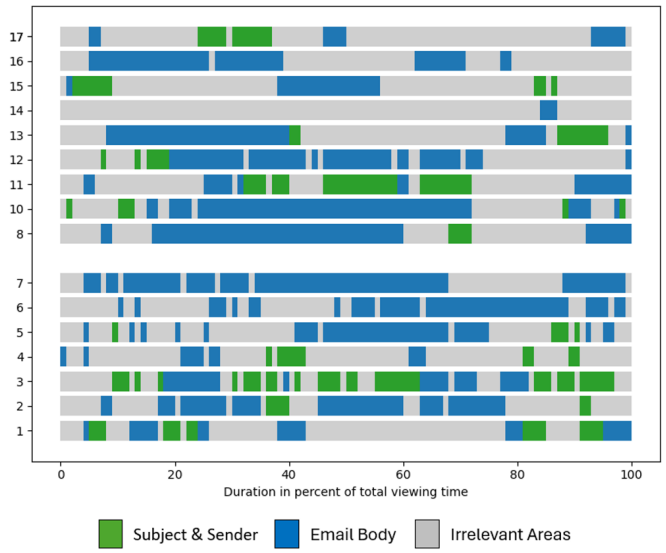

Figure 18. Scarf Plot for the Deutsche Bahn email: Visualizing AoI Transitions between phishing markers from participant using the PAS. Participants 1 to 7 are within the $Q_{0.05}$ and participants 8 to 16 are within the $Q_{0.95}$ of correctly identified emails. AoI hits from the PAS and the email itself are combined.

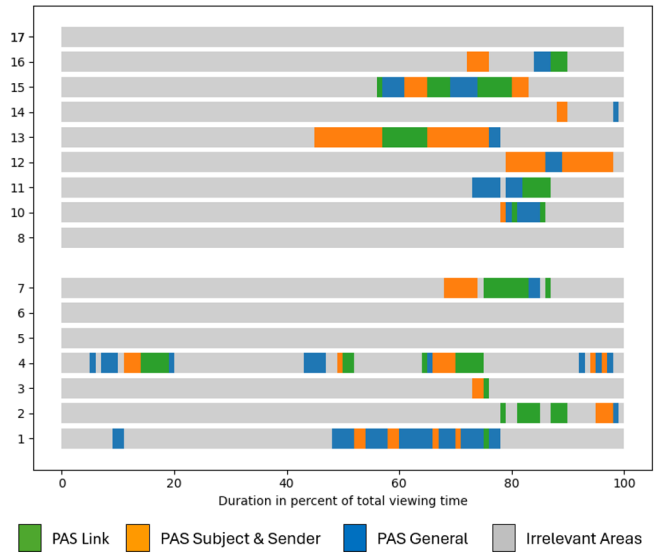

Figure 19. Scarf Plot for the Deutsche Bahn email: Visualizing AoI Transitions within the PAS. Participants 1 to 7 are within the $Q_{0.05}$ and participants 8 to 16 are within the $Q_{0.95}$ of correctly identified emails.

Concluding this, RQ7 can be answered: gaze patterns do differ between individuals who correctly identified all phishing emails and those who misclassified more to some extent. However, these differences are not uniform but manifest in multiple ways. First of all, participants who misclassified more emails tend to have shorter, more abrupt AoI viewing patterns, frequently switching between AoIs, particularly in the group with the PAS. Secondly, PAS usage varies based on phishing email complexity. When phishing markers were less

obvious, participants who correctly identified all emails were more likely to use the PAS toward the end of their decision-making process. Adding to this, for easier-to-detect phishing emails, participants who sorted all emails correctly studied the email body more carefully, suggesting they were quick to identify obvious phishing markers and validated their findings by examining additional cues.

A. Summary of results

Demographic factors such as age, gender, and highest general education degree were found to have no significant impact on phishing recognition rates. However, working in different industry sectors and the number of weekly screen-based work hours had a notable effect on the time participants needed to recognize phishing attempts. Employees from the IT sector were able to recognize phishing emails much faster compared to those from the construction or utilities sectors. Interestingly, prior knowledge of the sender before the study did not influence participants' ability to identify phishing emails.

Grouping the participants into IT security experts and novices revealed that participants with higher IT security knowledge were significantly faster at recognizing phishing attempts. However, this did not translate to a higher accuracy in identifying phishing emails. This finding partially aligns with the results of Ribeiro et al. [11], where users were more likely to misclassify legitimate emails than phishing ones. However, unlike the results reported in [11], participants in this study spent significantly more time examining the sender area in legitimate emails, while phishing emails did not garner as much attention in this area.

Further analysis of Areas of Interest (AoI) hits revealed that participants who misclassified more phishing emails tended to glance at relevant phishing markers for shorter, more abrupt periods. In contrast, participants who correctly identified phishing emails appeared to engage with the AoIs more thoroughly, especially when the phishing email was well-made. Notably, AoI hits on the PAS indicated that participants who successfully identified phishing emails relied on the PAS primarily when the email was particularly convincing and typically just before making their final decision.

VI. Usability analysis

The previous paper already highlighted a positive correlation between high usability ratings of the PAS and participants' ability to correctly identify phishing emails [1]. Combining these findings with insights from RQ7, the question arises: Do usability ratings of the PAS also correlate with the users' gaze patterns, particularly when interacting with the PAS itself? Previous research in fields such as machine learning and human-computer interaction has shown that specific eye movement patterns can reflect the usability of a system [31] [16]. This observation leads to the introduction of the last research question 8: Does the usability (and possibly the user experience) of the PAS relate to participants' gaze patterns during their?

For RQ8, only participants who interacted with the PAS will be considered. Furthermore, since this analysis is not tied to specific email stimuli, gaze patterns from all emails in the study will be aggregated into a single timeline and analyzed as a whole.

1) Usability and UX Questionnaires: Two types of questionnaires were used to assess usability and user experience (UX): the System Usability Scale (SUS) and the short version of the User Experience Questionnaire (UEQ-S).

The SUS measures perceived system usability through a ten-item questionnaire. Developed by John Brooke in the late 1990s, the SUS was designed to align with the ISO 9241-110 standard, making it universally applicable across different systems and contexts [32]. Respondents answer on a five-point Likert scale, with half of the items formulated to elicit agreement and the other half to elicit disagreement. The final usability score is calculated by weighing the responses, yielding a score between 0 and 100, with higher scores indicating better usability. In some cases, this score is further categorized into grades from A to F, with a score around 50 or lower indicating poor usability [33].

In contrast, the UEQ assesses not only usability but also the overall user experience (UX). Developed by Laugwitz et al., the UEQ distinguishes between pragmatic quality (associated with usability) and hedonic quality (related to UX) [34]. It's important to note that these two dimensions represent one of many perspectives on usability and UX, with alternative definitions existing in the literature [35]. The UEQ originally consists of 26 items; however, since evaluating the PAS's usability and UX was not the primary focus of this study, only the short version - UEQ-S - was used. This version contains eight items, offering a concise but comprehensive assessment of both pragmatic and hedonic quality [36]. Like the SUS, it uses a Likert scale, but with seven points instead of five. Scores for both pragmatic and hedonic quality are calculated by averaging the responses to the relevant items, with scores below 3.2 indicating poor results and those above 4.8 indicating high results [37].

2) Usability and UX Results: Similar to RQ7, percentiles are employed to categorize individuals into groups representing opposite extremes on the usability scale. Yet, in this case, the $Q_{0.10}$ and $Q_{0.90}$ percentiles are used, as the usability and UX scores fluctuated more than the number of correctly identified phishing emails. This leads to the following percentiles:

- SUS $Q_{0.1} = 52.75$ ($n = 6$)
- SUS $Q_{0.9} = 92.5$ ($n = 7$)
- UEQ-S Pragmatic Quality $Q_{0.1} = 4$ ($n = 8$)
- UEQ-S Pragmatic Quality $Q_{0.9} = 7$ ($n = 10$)
- UEQ-S Hedonic Quality $Q_{0.1} = 3.25$ ($n = 7$)
- UEQ-S Hedonic Quality $Q_{0.9} = 6.25$ ($n = 7$)

Starting with usability, an analysis of PAS usage — measured by any Area of Interest (AoI) hits within the sidebar — reveals a significant difference in behavior between participants who rated the tool as less usable and those who rated it as highly usable. Participants who perceived the PAS as less usable (SUS Score ≤ 52.75) used the tool significantly

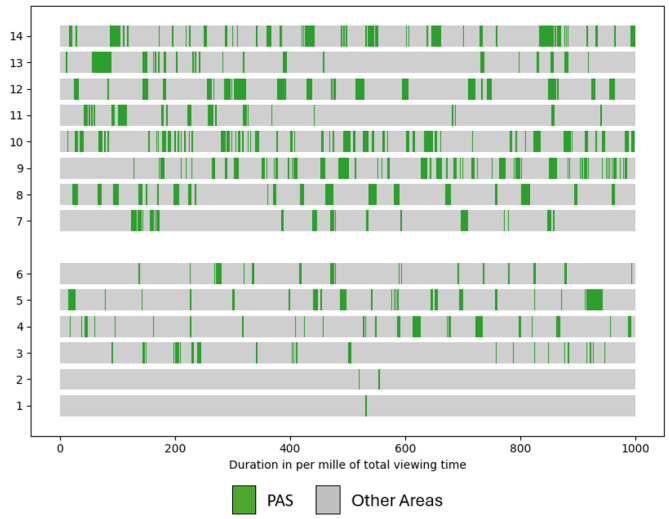

Figure 20. Scarf plot visualizing AoI hits within the PAS in relation to the SUS score. Participants 1 to 6 are within the $Q_{0.1}$ and participants 7 to 14 are within the $Q_{0.9}$ the SUS score.

less (see Figure 20). In the scarf plots, green areas indicate PAS usage, while grey areas represent time spent looking at the email itself or other PAS-unrelated screen areas. This viewing behavior was expected, as both low usability and user experience typically leads to reduced acceptance and adoption of software [38].

In contrast, participants who rated the PAS as highly usable (SUS Score ≥ 92.5) tended to use the tool more frequently and for longer periods. A similar pattern emerges when examining the pragmatic quality results of the UEQ-S (see Figure 21). Those who rated the PAS's pragmatic quality as low to neutral (UEQ-S Pragmatic Quality Score ≤ 4) also used the tool less. However, one outlier - participant 2 - used the tool just as much as participants who rated the pragmatic quality as high (UEQ-S Pragmatic Quality Score $= 7$).

Figure 21. Scarf plot visualizing AoI hits within the PAS in relation to the UEQ Pragmatic Quality score. Participants 1 to 6 are within the $Q_{0.1}$ and participants 7 to 14 are within the $Q_{0.9}$ the UEQ Pragmatic Quality score.

Interestingly, the results for hedonic quality show an unexpected trend. As seen in Figure 22, participants who rated the hedonic quality as low to neutral (UEQ-S Hedonic Quality Score ≤ 3.25) actually spent more time looking at the tool than those who rated it as high (UEQ-S Hedonic Quality Score ≥ 6.25). This could indicate confusion or a lack of trust in the PAS, leading to prolonged examination of the tool. Alternatively, participants proficient in detecting phishing emails may generally need less time overall, and their efficiency leads to them spending less time with the PAS. In a study setting where participants are primed and racing against the clock, this seems plausible — there is little time to appreciate the design, while visual irregularities may cause the participant to stop. However, without additional UX data, this cannot be explained definitively.

Figure 22. Scarf plot visualizing AoI hits within the PAS in relation to the UEQ Hedonic Quality score. Participants 1 to 6 are within the $Q_{0.1}$ and participants 7 to 14 are within the $Q_{0.9}$ the UEQ Hedonic Quality score.

Despite these contradicting findings, all three scarf plots clearly demonstrate that gaze patterns vary between individuals with differing usability and UX ratings. Nevertheless, hypothesis H8 must be partially rejected: While higher usability did indeed lead to increased PAS usage, participants who perceived UX as high actually spent less time looking at the tool.

VII. Limitations

It is a consistent pattern throughout the entire study that differences between groups were only observed in the amount of time needed to complete the email classification task, but never in the number of correctly identified emails. This could be attributed to the nature of statistical tests: many tied values within a recorded quantity reduce the statistical power to detect significant differences. As a result, while participants may vary in the speed at, which they complete the task, their accuracy appears to remain consistently high across all groups. This finding emphasizes the importance of measuring multiple dimensions of performance when evaluating differences between

groups. Focusing solely on one aspect, such as accuracy, may overlook meaningful variations in other areas, such as task efficiency.

VIII. Conclusion and Future Work

The findings from this study indicate that prior training or specific knowledge, such as being an IT security expert, do not influence the number of correctly identified phishing emails. This suggests that errors in identifying phishing emails are more likely because of genuine user mistakes and oversights, rather than a lack of knowledge. Despite this, it was observed that certain factors, such as education level, industry sector, IT security knowledge, and weekly screen-based work hours, had a significant impact on the time required to recognize phishing attempts. Participants with higher education levels, more IT security knowledge, greater weekly screen exposure, or those working in the IT sector, performed the task more efficiently, needing less time to identify phishing emails. Interestingly, no significant differences were found between individuals with IT-related backgrounds and those without, indicating that phishing detection training may be beneficial for all participants, regardless of their profession or expertise.

The nature of statistical testing, particularly with discrete variables like the number of correctly identified emails, makes it difficult to detect significant differences between groups when there are many tied values. While it is challenging to present participants with large datasets due to time constrains, especially demographic correlation analysis would benefit from a bigger dataset. With more data it could even be possible to measure influences of demographic factors, which yielded no effect in this study. However, this data limitation lies in the nature of eye-tracking studies, which are not infinitely scalable due to the need for specific technical equipment and participant monitoring by the conductors.

Future research could explore ways to improve phishing detection across all user groups, including those with limited IT security knowledge. Further studies could investigate whether longer or more detailed training sessions can enhance detection accuracy and speed for participants with less prior knowledge. Additionally, expanding the study to include larger and more varied dataset, perhaps with more frequent exposure to phishing attempts or even a redesigned version of the PAS, would help address the limitations of the current approach and provide further insights into the role of experience and training in phishing recognition.

Acknowledgments

This study was conducted as part of the EU-funded EDIH *Digital Innovation Ostbayern (DInO)*. DInO is funded by the European Union (Project Reference 101083427) and the European Funds for Regional Development (EFRE) (Project Reference 20-3092.10-THD-105). The eye-tracking study was approved by the Joint Ethics Committee of the Bavarian Universities (GEHBa) with the reference number GEHBa-202312-V-155-R.

This article and the one presented in [1] are the result of a joint collaboration between all authors, with each contributing equal effort. Therefore, authorship is alternated in this paper, which does not reflect a change in contribution levels.

The individuals in Figure 4 - despite being only shown from behind - both agreed to be shown in this paper.

Data

The eye-tracking and questionnaire data collected and evaluated in this study is free to use and can be found on Zenodo under the following link doi.org/10.5281/zenodo.13171791.

References

[1] M. Kristen, F. Engl, and J. Mottok, "Enhancing phishing detection: An eye-tracking study on user interaction and oversights in phishing emails," in *SECURWARE 2024, The Eighteenth International Conference on Emerging Security Information, Systems and Technologies*, 2024.

[2] Bundesamt für Sicherheit in der Informationstechnik (BSI), "Die Lage der IT-Sicherheit in Deutschland 2024," de, 2022.

[3] M. Bada, A. M. Sasse, and J. R. Nurse, "Cyber security awareness campaigns: Why do they fail to change behaviour?" *arXiv preprint arXiv:1901.02672*, 2019.

[4] A. M. Sasse, S. Brostoff, and D. Weirich, "Transforming the 'weakest link' — a human/computer interaction approach to usable and effective security," *BT Technology Journal*, vol. 19, no. 3, pp. 122–131, Jul. 1, 2001. DOI: 10.1023/A:1011902718709.

[5] A. Heinemann and G. Schembre, "Zur Wirksamkeit von Security Awareness Maßnahmen," ger, in *DACH Security Tagungsband 2017: Bestandsaufnahme, Konzepte, Anwendungen, Perspektiven*, P. Schartner and A. Baumann, Eds., Klagenfurt (Österreich): Alpen-Adria-Universität, 2017, ISBN: 978-3-00-057290-6.

[6] D. Lain, K. Kostiainen, and S. Capkun, "Phishing in organizations: Findings from a large-scale and long-term study," *2022 IEEE Symposium on Security and Privacy (SP)*, p. 9, 2022.

[7] H. Abroshan, J. Devos, G. Poels, and E. Laermans, "Phishing happens beyond technology: The effects of human behaviors and demographics on each step of a phishing process," *IEEE Access*, vol. 9, pp. 44 928–44 949, 2021. DOI: 10.1109/ACCESS.2021.3066383.

[8] A. Darwish, A. E. Zarka, and F. Aloul, "Towards understanding phishing victims' profile," in *2012 International Conference on Computer Systems and Industrial Informatics*, 2012, pp. 1–5. DOI: 10.1109/ICCSII.2012.6454454.

[9] A. K. Ghazi-Tehrani and H. N. Pontell, "Phishing evolves: Analyzing the enduring cybercrime," *Victims & Offenders*, vol. 16, no. 3, pp. 316–342, 2021. DOI: 10.1080/15564886.2020.1829224.

[10] J. L. Orquin and K. Holmqvist, "Threats to the validity of eye-movement research in psychology," *Behavior Research Methods*, vol. 50, no. 4, pp. 1645–1656, Aug. 2018. DOI: 10.3758/s13428-017-0998-z.

[11] L. Ribeiro, I. Guedes, and C. Cardoso, "Eyes on phishing emails: An eye-tracking study," *Journal of Experimental Criminology*, 2024.

[12] G. A. Thomopoulos, D. P. Lyras, and C. A. Fidas, "A systematic review and research challenges on phishing cyberattacks from an electroencephalography and gaze-based perspective," *Personal and Ubiquitous Computing*, 2024.

[13] D. Baltuttis and T. Teubner, "Effects of visual risk indicators on phishing detection behavior: An eye-tracking experiment," *Computers & Security*, vol. 144, p. 103 940, 2024. DOI: 10.1016/j.cose.2024.103940.

[14] S. Zhuo, R. Biddle, J. Daniel Recomendable, G. Russello, and D. Lottridge, "Eyes on the phish(er): Towards understanding users' email processing pattern and mental models in phishing detection," in *Proceedings of the 2024 European Symposium on Usable Security*, ser. EuroUSEC '24, Association for Computing Machinery, 2024, pp. 15–29, ISBN: 9798400717963. DOI: 10.1145/3688459.3688465.

[15] L. Zhou, J. Lim, and D. Zhang, "The effects of email illegitimacy and phishing behavior training on eye movement behavior in spear phishing detection," in *Proceedings of the 57th Hawaii International Conference on System Sciences*, 2024.

[16] B. Xing *et al.*, "User-attention based product aesthetics evaluation with image and eye-tracking fusion data analysis," in *2023 15th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Aug. 2023, pp. 84–87.

[17] P. Sulikowski and T. Zdziebko, "Deep learning-enhanced framework for performance evaluation of a recommending interface with varied recommendation position and intensity based on eye-tracking equipment data processing," *Electronics*, vol. 9, no. 2, p. 266, Feb. 2020, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[18] M. C. Sáiz-Manzanares *et al.*, "Analysis of the learning process through eye tracking technology and feature selection techniques," *Applied Sciences*, vol. 11, no. 13, p. 6157, Jan. 2021, Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

[19] T. Ezer, M. Greiner, L. Grabinger, F. Hauser, and J. Mottok, "Eye tracking as technology in education: Data quality analysis and improvements," in *ICERI2023 Proceedings*, ser. 16th annual International Conference of Education, Research and Innovation, Seville, Spain: IATED, Nov. 2023, pp. 4500–4509, ISBN: 978-84-09-55942-8. DOI: 10.21125/iceri.2023.1127.

[20] G. Kaiser, *Marktanteile der Top 10 E-Mail-Clients weltweit im März 2025*, https://de.statista.com/statistik/daten/studie/688163/umfrage/marktanteile-der-e-mail-clients-weltweit/, Accessed: 2025-05-19.

[21] C. Blake, "Eye-Tracking: Grundlagen und Anwendungsfelder," ger, in *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft*, W. Möhring and D. Schlütz, Eds., Wiesbaden: Springer Fachmedien Wiesbaden, 2013, pp. 367–387, ISBN: 978-3-531-18776-1. DOI: 10.1007/978-3-531-18776-1_20.

[22] M. Burch and K. Kurzhals, "Teaching eye tracking: Challenges and perspectives," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. ETRA, pp. 1–17, 2024.

[23] B. T. Carter and S. G. Luke, "Best practices in eye tracking research," *International Journal of Psychophysiology*, vol. 155, pp. 49–62, 2020.

[24] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, Dec. 1965. DOI: 10.1093/biomet/52.3-4.591. eprint: https://academic.oup.com/biomet/article-pdf/52/3-4/591/962907/52-3-4-591.pdf.

[25] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[26] H. Abdi, "The bonferonni and šidák corrections for multiple comparisons," *Encyclopedia of Measurement and Statistics*, vol. 3, Jan. 2007.

[27] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[28] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, Jun. 1938. DOI: 10.1093/biomet/30.1-2.81. eprint: https://academic.oup.com/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf.

[29] W. G. Cochran, "The χ^2 Test of Goodness of Fit," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 315–345, 1952. DOI: 10.1214/aoms/1177729380.

[30] C.-K. Yang and C. Wacharamanotham, "Alpscarf: Augmenting scarf plots for exploring temporal gaze patterns," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '18, Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–6, ISBN: 9781450356213. DOI: 10.1145/3170427.3188490.

[31] P. Wang, H. Yang, J. Hou, and Q. Li, "A machine learning approach to primacy-peak-recency effect-based satisfaction prediction," *Information Processing & Management*, vol. 60, no. 2, p. 103 196, Mar. 1, 2023.

[32] J. Brooke, "SUS: A 'quick and dirty' usability scale," *Usability Evaluation in Industry*, vol. 1, pp. 189–194, 1996.

[33] R. A. Grier, A. Bangor, P. Kortum, and S. C. Peres, "The system usability scale: Beyond standard usability testing," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 57, no. 1, pp. 187–191, 2013. DOI: 10.1177/1541931213571042.

[34] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4*, Springer, 2008, pp. 63–76.

[35] M. Richter and M. D. Flückiger, *Usability und UX kompakt* (IT kompakt). Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.

[36] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Design and evaluation of a short version of the user experience questionnaire (UEQ-S)," *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.*, 2017.

[37] S. Martin, "User experience questionnaire handbook," de, 2023.

[38] M. Hassenzahl, "The effect of perceived hedonic quality on product appealingness," *International Journal of Human-Computer Interaction*, vol. 13, no. 4, pp. 481–499, 2001.