Lending a Hand: The Effectiveness of Support Systems in Assisting Users to Detect Phishing Attacks

Katharina Schiller
katharina.schiller@hof-university.de
Hof University of Applied Sciences
Hof, Bavaria, Germany

Jörg Scheidt
joerg.scheidt@hof-university.de
Hof University of Applied Sciences
Hof, Bavaria, Germany

Florian Adamsky
florian.adamsky@hof-university.de
Hof University of Applied Sciences
Hof, Bavaria, Germany

Zinaida Benenson
zinaida.benenson@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Erlangen, Bavaria, Germany

Abstract

We investigate the effectiveness of anti-phishing support systems through a quantitative study involving 453 participants. To this end, we developed a tool that allows participants to immerse themselves in a realistic setting, tasked with classifying emails as either phishing or legitimate, while being assisted by support systems. Despite the prevalence of support systems in webmailers and email clients, our results indicate no significant difference in correctly assessing emails of varying difficulty between these systems and the control group. We found a minor negative effect of the support system that uses tooltips compared to other support systems. In the subsequent survey, we found that the support systems are appreciated and considered helpful by users, as supported by the results of the UEQ-S, even if they have no observable effect. Email context, such as the contact list, as well as hovering over the links, had stronger effects on the classification than the tested support systems.

CCS Concepts

• **Security and privacy → Usability in security and privacy**; • **Social and professional topics → Phishing**; • **Human-centered computing → User studies**.

Keywords

phishing, anti-phishing support system, effectiveness, tooltip, external marker, banner

ACM Reference Format:
Katharina Schiller, Jörg Scheidt, Florian Adamsky, and Zinaida Benenson. 2026. Lending a Hand: The Effectiveness of Support Systems in Assisting Users to Detect Phishing Attacks. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3772318.3791490

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791490

1 Introduction

Not a day goes by without us receiving at least one email designed to trick us into clicking on malicious links or attachments, revealing our data on websites, or installing malware. These social engineering attacks are known as *phishing* and are among the most grave security threats to businesses. According to a recent report [1], in the first quarter of 2025, 41 % of all account compromises can be attributed to phishing attacks. These are not attempts, but confirmed cases where the user's identity was compromised.

To detect phishing emails, good advice used to be to look for spelling and grammatical errors in the email. However, generative AI and AI-based translation services are at everyone's fingertips. With such services, phishing campaigns that are difficult to recognize are easy to launch. As a result, phishing attacks become an even bigger problem with AI. Technical countermeasures such as DomainKeys Identified Mail (DKIM), Sender Policy Framework (SPF), and Domain-based Message Authentication Reporting and Conformance (DMARC) can filter suspicious emails and increase the effort required for an adversary. However, these are often difficult to implement and break existing workflows, such as mailing lists, and require high maintenance, including changing cryptographic keys periodically [28, 36].

As a result, we need additional client-side instruments that help users detect phishing emails. One widespread client-side instrument to detect phishing is a *support system*, which provides hints and highlights email elements that indicate phishing, and should help users detect suspicious emails more effectively. To the best of our knowledge, there are no studies that comparatively investigate the effectiveness of several support systems in a realistic scenario.

Research questions (RQs). We consider the following research questions:

RQ1 How effective are various support systems in helping users to detect phishing emails?
RQ2 Do support systems lead to overcautious behavior (misclassifying legitimate emails as phishing) compared to the control group?
RQ3 How do false positives (marking legitimate emails as phishing) and false negatives (not marking a phishing email as phishing) in the classification by support systems influence phishing detection by users?
RQ4 What is the perception of support systems by users?

To investigate these questions, we developed an interactive tool where participants slip into the role of an HR employee tasked with classifying emails as legitimate or phishing. We grouped 453 participants into five groups, which included the following support systems: external marker, TORPEDO tooltip [55], warning banner, spam label, and a control group with no support system. We also investigated whether participants can classify emails correctly even when the support system is incorrect. The scenario incorporates false positives, where our tool displays support systems for legitimate emails, and false negatives, in which a phishing email lacks a support system. The participants needed to classify 15 legitimate emails and 3 phishing emails and afterwards participated in a survey.

Contributions. Even though the biggest email providers such as Gmail and Outlook integrate support systems, we show that, in our experimental design, the tested support systems do not influence the correct classification of emails in comparison to the control group. However, we could find an increase in overcautious behavior (marking legitimate emails as phishing) for the tooltip, with participants performing nearly one-third worse than with the other support systems on specific emails.

Furthermore, our study shows that the context that our tool provides (especially the contact book, and to some extent the calendar) has more impact on the classification results than the actual support systems. Furthermore, one of the most important factors for correct classification remains hovering over the link.

Finally, the survey reveals that participants exhibited high trust in the support systems, with the majority reporting increased phishing awareness and perceiving the systems as helpful. This finding is corroborated by the User Experience Questionnaire's Short Version (UEQ-S) analysis, where all support systems achieved positive pragmatic quality scores above 1.0, indicating user satisfaction.

By analyzing user reactions to support systems, including classification errors (false positives and false negatives), we bridge the gap between user behavior and technical behavior. We show in which cases a redesign of anti-phishing systems would be necessary, rather than blaming the users for incorrect behavior driven by the systems.

2 Background and Related Work

The range of technologies available to prevent phishing emails from reaching recipients' inboxes in the first place is vast. Nevertheless, new attack strategies and concepts continue to emerge [2, 53], allowing phishing emails to reach users. Stricter technical phishing filters would result in legitimate emails not arriving either, leading to frustration among users. At the same time, users trust and rely on these security measures provided by their companies, which may lead to a false sense of security [16, 20, 21, 61]. Ultimately, this leaves the responsibility with the user, who must classify emails as legitimate or phishing. To assist users in their decision-making, several approaches are available. We follow Kumaraguru et al. [33] and divide related work into *training users* and *warning users.*

2.1 Training Users

Various studies have investigated the effectiveness of different training methods, including instructor-based training [14, 50, 54], text-based training [14, 37, 41, 47, 50, 54], group discussions [15, 41], and role-playing exercises [15]. Furthermore, Reinheimer et al. [39] investigated the frequency of training and recommend reminders to maintain permanent learning effects. Some approaches also utilize gamification and interactive elements [12, 23, 47, 59, 60]. Particularly, companies and organizations rely on the training of employees and members to protect the entire unit from phishing. A common and widely examined training method is embedded training [32]. Immediately after falling for a simulated phishing email and clicking on a link within it, users receive training in various scenarios. Regarding such phishing simulations, many studies focused on their effectiveness [19, 20, 29, 48, 62], user acceptance [42, 46] or used it to evaluate the effectiveness of other training methods [17, 37, 57]. While a range of studies [14, 19, 25, 29, 48, 62] see advantages over other training methods, some studies note negative side effects [11, 21, 34, 56, 61]. Reasons for limited effectiveness of phishing simulations include that this material is provided only to the limited group of users who clicked on a particular simulated phishing email, and even these users not reading through the training material provided after clicking [13, 26].

2.2 Warning Users (Support Systems)

There is a category of tools designed to warn users and raise awareness, classified as *support systems.* Unlike technical measures that hide emails from users, these systems aim to assist users in their decision-making.

Warnings Messages. Active warnings in comparison to passive ones can be effective [3, 18], but the frequency of warnings and the amount of false positives is the habituation effect [9, 31, 51]. To avoid this effect, various studies [7, 31, 38] adapted the design and analyzed different variations of warning messages. Lain et al. [34] compared verbose and succinct warnings and found no difference in their effectiveness. A recent study by Tolsdorf et al. [52] found that while warnings in phishing simulations are sufficient, their format does not impact effectiveness. However, none of the studies examined how false positives and false negatives affect warning effectiveness, as all simulated phishing messages either included or excluded warnings.

Labels. Many email systems assign classification labels to flag messages as spam or junk. EmailVeritas[1] provides *Phishing Detector*, a plugin for Outlook and Gmail, which has its own scoring system and adds colored labels: red for phishing, yellow for spam, and green for legitimate emails.

URL Parsing Support. Studies [4, 40] show that users often struggle to parse URLs correctly, which is an essential indicator of suspicious emails. Domain highlighting has been developed to overcome this problem, and one that is particularly interesting is Torpedo by Volkamer et al. [55] Alongside color-coded assessments and domain highlighting, the link is blocked for 3 seconds if the domain is unknown or suspicious, encouraging users to scrutinize the URL.

[1] https://www.emailveritas.com/

In their study, the tooltip improved phishing detection rates and increased users' confidence in their decisions. Additionally, Lain et al. [35] explored methods to make URLs more understandable, which involve user interaction and serve as a barrier to users before they access potentially malicious websites.

External Marker. Another widely used method in companies is to mark external emails accordingly. Marking external emails can be achieved by adding a [External] tag in the subject line or by using icons with explanations above the email content. Schiller et al. [42] show that this external marker was a relevant indicator to identify phishing for many participants, especially those with limited external contacts. However, Tolsdorf et al. [52] found no or inconsistent effects of the marker on susceptibility to simulated phishing attacks.

Other Tools and Factors. Zheng and Becker [63] conducted a formative evaluation to assess the usability of three self-developed email security tools using mockups. A warning banner with an assessment of the email and a button to check previous correspondence with the sender were considered helpful, while a notice that colleagues had marked an email as suspicious was rather confusing. Although the users referred to technical cues for evaluating emails, they created their own mental models. For instance, the relationship with the sender of the email had an important influence. In general, users pay attention to the content and context of the emails. Other studies [8, 22, 25, 42] also report these factors as primary cues users rely on to identify phishing. At the same time, users also contribute their former experiences to their assessments, or use the real world to confirm their assumption [58].

2.3 Comparative Evaluations of Anti-Phishing Support Systems and Research Gap

Previous studies tended to examine support systems of one type, such as different variants of SSL warnings [3], assistance in correctly reading URLs [35] or warning messages in several forms [34, 38]. Zheng and Becker [63] qualitatively investigated several self-designed support system variants that are not integrated into real email clients. Tolsdorf et al. [52] investigated the effectiveness of various realistic anti-phishing support systems, including different types of warnings and the widely used marking of emails as external, in a large university hospital. Active warnings were the most effective, while the external marker had little to no effect.

Our study investigates four support systems that exist in practice within a fictional, controlled scenario. Implementing all these support systems and testing them in a field study similar to Tolsdorf et al. [52] would have meant a very high (to the point of prohibitive) effort for the involved organizations. Although we do not exclude the possibility of implementing a similar study in the field, we decided to take the first step in a simulated controlled environment. To simulate a real-world scenario, we used a ratio of legitimate emails and phishing emails that is closer to reality than many previous settings, with varying degrees of classification difficulty, which we tested in a preliminary study.

Furthermore, we collect participants' subjective opinions on the support systems, measure the user experience by using the User Experience Questionnaire (UEQ) and compare these opinions and measurements across support systems.

Most important difference of our study to all previous studies is the inclusion into the investigation false positives and false negatives in the classification by support systems, which is close to reality. Indeed, if it was possible to develop support systems that never make mistakes, we would not need support systems. In this case, all phishing emails would have been filtered out by a flawless mechanism. In contrast to previous studies, our controlled setting allows us to include false positives and false negatives to examine their effects. We also investigate whether support systems may lead to overcautious behavior, i. e., to participants mistakenly classify more legitimate emails as phishing.

3 Methodology

This section describes the methodology for the quantitative and qualitative parts of our study. Figure 1 shows our general approach to the study design. The tool begins with a scenario introduction, followed by a tutorial on its functionality, which participants must complete. Then, we divided the 453 participants into five groups, including the four support systems and one control group. Each group needed to classify 18 emails in random order, including 3 phishing emails and 15 legitimate ones. After that, all groups filled out a general survey about which cues they thought were most important for classifying phishing. Then, we asked only the support system groups if they had recognized the support system. If so, we presented a survey about the support system; otherwise, we redirected them to the demographic questionnaires. Since the control group did not see a support system, we redirected this group directly to the demographic questionnaires.

3.1 Interactive Tool

The objective was to create a realistic environment and scenario that allowed participants to immerse themselves in the situation. Additionally, some support systems, such as the tooltip, require a certain amount of interaction to be noticed. Participants should be able to interact with the emails to a certain extent, but not be able to leave the system. We decided to develop a tool inspired by Outlook but simplified, and to integrate emails, user context, and support system into it.

3.1.1 *Study Scenario.* User context plays a particularly influential role in recognizing phishing emails [8, 22, 25, 42, 49]. Therefore, we place great importance on the context in our study. The participants slip into the role of *Alex Schulz*, an employee of the HR department of a fictional IT company *Smartcompany*. This name is gender-neutral in Germany; we chose it so that both women and men can identify with it. Alex Schulz has the task to classify 18 emails as fraudulent or legitimate, while being assisted by a support system (if not being in the control group). In addition to the email classification, the tool provides a calendar, a contact book, and additional information for the context.

3.1.2 *Tool.* Figure 2 displays multiple screenshots of the user interface of our self-developed experimental tool. The source code for our tool is open source and available via Git[2]. The tool and

[2] https://github.com/iisys-sns/phishing-study

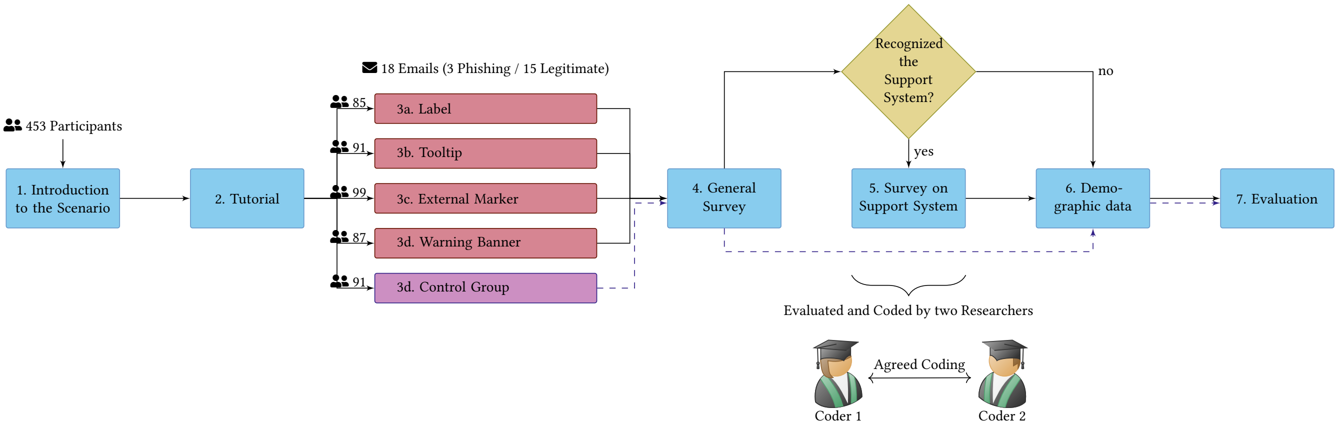


Figure 1: Study design: The blue boxes represent work steps, and the red boxes indicate the different groups where our tool showed support systems. The purple box represents the control group without a support system. All groups evaluated 18 emails in random order, including 3 phishing emails and 15 legitimate ones. The numbers in front of the groups denote the number of participants in this group. The dashed purple line indicates the different path of the control group through our study. The survey was evaluated and coded by two independent researchers, first separately and then in multiple meetings together.

the survey were tested several times, including with older people. We made adjustments to the tool and survey after every iteration based on the feedback and observed issues. In the process, we considered usability, functionality, the valid recording of results and completion time.

User Interface. The interface of our tool displays one email at a time in random order in the center. For each displayed email, a green *legitimate* and a red *fraudulent* button are shown, allowing the participant to classify the email. There is a counter at the top right displayed in Figure 2a that visualizes how many emails still need to be processed. On the left-hand side of the tool, several non-clickable icons are displayed for applications that are used within the company. At the bottom, there are icons for a calendar, a contact book, and a job advertisement. These open up and provide further information if clicked.

Calendar, Contacts, and Additional Information. The calendar in Figure 2b highlights the current date with a blue circle. Furthermore, the calendar highlights two recent appointments in red—an interview with a potential job candidate and a company event. The calendar highlights two other relevant items in yellow: a technical delivery and the expiration of the password. The latter calendar entry was to show that this is a valid process in the company.

The contact book in Figure 2b shows several contacts in the company who are familiar to the fictional character. In addition to the name, email address, and, if available, telephone number, the department is also listed in brackets. By clicking on the last icon in the menu bar below, the user will see a schematic representation of the job offer, indicating that SmartCompany is currently looking for new employees. The participant's task is to assume the role of Alex Schulz and classifying emails in the inbox as phishing or legitimate.

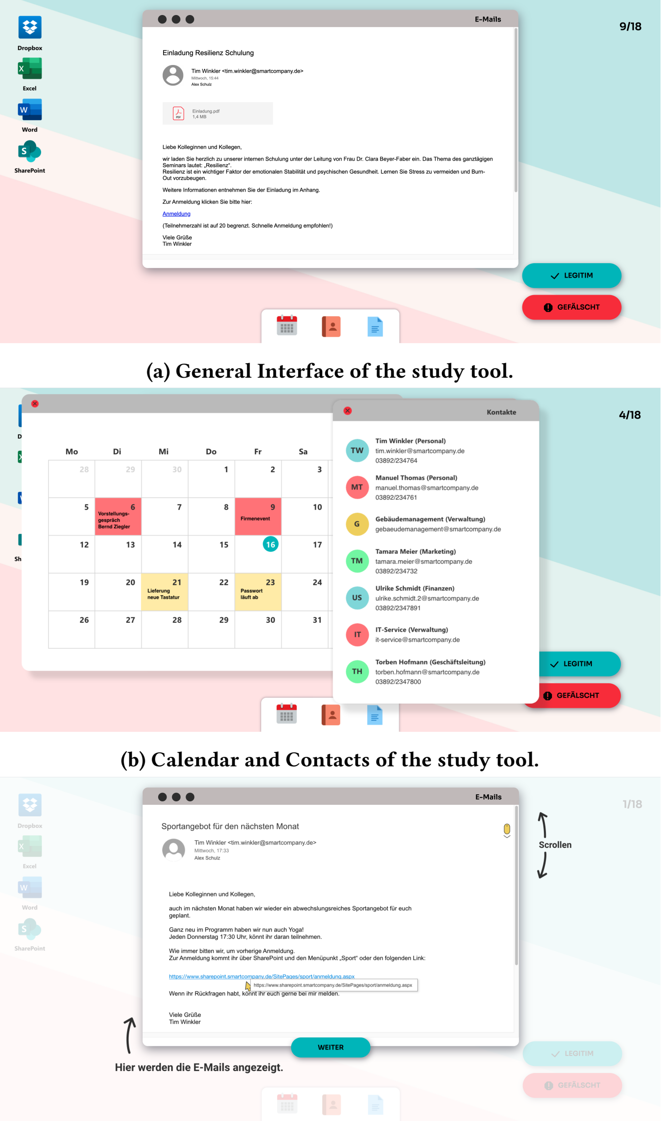
Tutorial. Before participants can begin the study, they must complete a tutorial that explains the tool's full functionality. Different animations demonstrate that participants can scroll through emails and hover over links to reveal the entire URL. The latter is not explicitly explained in text form to avoid guiding participants who do not usually hover over links in this direction. The tutorial explains in text form and with illustrated arrows that the icons on the left-hand side are not clickable; however, SmartCompany utilizes these applications. After that, participants can continue by clicking a blue button labeled *next*, and then the tool prompts them to open the calendar.

The tutorial only continues if the participants actually click on the calendar to ensure that they are aware of the functionality and the information inside. Next, the tutorial indicates that the other icons in the toolbar are also clickable. However, we do not require participants to open the contact book or the job offer to keep the tutorial brief. The final step in the tutorial is to classify an email about the company's sports program as phishing or legitimate using the buttons at the bottom right. An animation alternately shows that the two buttons are clickable. The participants receive information on whether their classification was correct or incorrect. However, this is only the case for the tutorial; in the main study, we show the results of the evaluation after completion.

Behavioral Data. Besides the final classification of the email by the user, the tool also captures various behavioral data. Firstly, this is the time a participant spends on an email. Secondly, it records when a user hovers over a link with the mouse. It should be noted that this can also happen accidentally when reading the email. Thirdly, it tracks whether a user has accessed the context elements: calendar, contacts, and job offer in an email.

3.2 Support Systems

We analyzed popular webmailers (e. g., Gmail, Outlook Online), email clients (e. g., Mozilla Thunderbird, Outlook), and plugins for these and searched the academic literature for support systems. We found three support systems used in the real world: external



Einladung Resilienz Schulung

Tim Winkler <tim.winkler@smartcompany.de>
Mittwoch, 15:44
Alex Schulz

Einladung.pdf
1,4 MB

Liebe Kolleginnen und Kollegen,

wir laden Sie herzlich zu unserer internen Schulung unter der Leitung von Frau Dr. Clara Beyer-Faber ein. Das Thema des ganztägigen Seminars lautet: „Resilienz".
Resilienz ist ein wichtiger Faktor der emotionalen Stabilität und psychischen Gesundheit. Lernen Sie Stress zu vermeiden und Burn-Out vorzubeugen.

Weitere Informationen entnehmen Sie der Einladung im Anhang.

Zur Anmeldung klicken Sie bitte hier:

Anmeldung

(Teilnehmerzahl ist auf 20 begrenzt. Schnelle Anmeldung empfohlen!)

Viele Grüße
Tim Winkler

(a) General Interface of the study tool.

Mo	Di	Mi	Do	Fr	Sa
28	29	30	1	2	3
5	6 Vorstellungsgespräch Bernd Ziegler	7	8	9 Firmenevent	10
12	13	14	15	16	17
19	20	21 Lieferung neue Tastatur	22	23 Passwort läuft ab	24
26	27	28	29	30	31

TW Tim Winkler (Personal)
tim.winkler@smartcompany.de
03892/234764

MT Manuel Thomas (Personal)
manuel.thomas@smartcompany.de
03892/234761

G Gebäudemanagement (Verwaltung)
gebaeudemanagement@smartcompany.de

TM Tamara Meier (Marketing)
tamara.meier@smartcompany.de
03892/234732

US Ulrike Schmidt (Finanzen)
ulrike.schmidt.2@smartcompany.de
03892/2347891

IT IT-Service (Verwaltung)
it-service@smartcompany.de

TH Torben Hofmann (Geschäftsleitung)
torben.hofmann@smartcompany.de
03892/2347800

(b) Calendar and Contacts of the study tool.

Sportangebot für den nächsten Monat

Tim Winkler <tim.winkler@smartcompany.de>
Mittwoch, 17:33
Alex Schulz

Liebe Kolleginnen und Kollegen,

auch im nächsten Monat haben wir wieder ein abwechslungsreiches Sportangebot für euch geplant.

Ganz neu im Programm haben wir nun auch Yoga!
Jeden Donnerstag 17:30 Uhr, könnt ihr daran teilnehmen.

Wie immer bitten wir, um vorherige Anmeldung.
Zur Anmeldung kommt ihr über SharePoint und den Menüpunkt „Sport" oder den folgenden Link:

https://www.sharepoint.smartcompany.de/SitePages/sport/anmeldung.aspx
https://www.sharepoint.smartcompany.de/SitePages/sport/anmeldung.aspx

Wenn ihr Rückfragen habt, könnt ihr euch gerne bei mir melden.

Viele Grüße
Tim Winkler

(c) Tutorial of the study tool.

Figure 2: Screenshots of our self-developed study tool. Subfigure (a) shows the general user interface of the tool, in which an email is displayed that the participant needs to classify as legitimate or fraudulent. Subfigure (b) shows the calendar and the menu with contacts, which the participant can open at any time during the study. Subfigure (c) shows the tutorial that every participant must complete to understand all functionality of the tool.

marker, banner, and label; from the academic literature, we found Torpedo by Volkamer et al. [55].

Our experimental tool, described in Section 3.1, is not a real email client. Therefore, we needed to imitate the functionality of the tested support systems in the tool. The support systems use various technical bases in the background, which are partially black boxes. To determine the functionality, we employed various strategies to replicate them as accurately as possible, e. g., sending or uploading some selected emails. We could not test the support systems in realistic conditions since we were unable to send emails from authentic phishing servers. Simultaneously, we had to assume that support systems can also be wrong, and that both false positives and false negatives can occur. This can happen, for instance, if an adversary has taken over an email account and uses it to send phishing emails, or if a phishing attempt is novel and still unknown to block listing platforms. However, our study does not focus on the functionality and performance of these support systems, but rather on whether they help to classify emails correctly. Next, we outline each support system and its integration into the experimental tool.

3.2.1 *External Marker.*

An external marker is a simple support system that shows if the email comes from an external server, i. e., not sent from an internal account. The marking of external emails is a well-known practice in business and must be configured organization-wide by the IT department. The email server of the recipient analyzes the header fields from the email and marks external emails by adding the external tag in front of the subject. In addition, various customizable HTML elements can be inserted above the email content. Schiller et al. [42] showed that the external markers are rated positively, and the participants see them as helpful in recognizing suspicious emails. Tolsdorf et al. [52] showed, however, that the external marker had no effect in recognizing phishing simulations.

For this study, we decided to place the tag `[EXTERNAL]` in front of the subject of the email and, additionally, above the content, an icon whith a short explanation that repeats the sender's address and states that the sender is from outside the organization. Figure 3a shows an example. The external marker is displayed for five of the emails, including two phishing emails. According to our scenario, these emails were sent from external senders. The external marker is based on a deterministic filtering mechanism that, when appropriately configured, cannot produce false negatives or false positives on its own. However, these false results can still happen. For example, a false positive would occur when an employee uses a private email account, leading the email client to label the emails as external. This is a special case that we did not consider in our study design. A false negative would be a compromised email account used for further phishing or spam attacks; these emails would not be labeled as external. We included a false negative in our study design, since this is a common approach for modern cyberattacks such as Emotet.

3.2.2 *Warning Banner.*

For the warning banner, we opted for a variant similar to the one used by Gmail. Emails with SPFs, DKIMs, or DMARCs errors are marked with a yellow warning banner above the email content, which states *"Be Careful With This Message"* and that Gmail could not verify the sender. The sender's address is also displayed again. There are also buttons on the banner that allow users to report the email as phishing or classify it as safe. We did not integrate these buttons into our study, as they take up additional space and are already available in a similar form at the bottom. Similar to the external marker, the warning banner is based on a deterministic mechanism. A false positive case would be a misconfigured email server that causes the receiving mail server to detect

SPF, DKIM, or DMARC errors. This is a special case that we did not consider in our study design. However, we included a false positive case: our difficult phishing email originating from a compromised internal account, i. e., the warning banner is displayed only in two of the three phishing emails. A screenshot of the warning banner can be seen in Figure 3d.

3.2.3 Spam Label. For the label, we focused on a plugin from EmailVeritas[3] called *Phishing Detector*, available for Outlook and Gmail. It uses the built-in functionality of the email client or webmail to add labels to emails, sorting them into categories. Upon activation, the plugin analyzes emails and labels them as phishing (red), spam (yellow), or legitimate (green). When a user enables the plugin, a sidebar opens, providing further information about the email and its categorization. At the same time, it provides guidance on handling suspicious emails. In our tool, the participants can open the sidebar by clicking on the label itself to get more information. Furthermore, users can view a summary of the threat highlights, containing information from the email header, such as the `FROM` or `Reply-To` fields, the creation date, or the user agent. Additionally, it displays a map that shows the location of the IP address from which the email originates. A screenshot can be seen in Figure 3c.

One problem with the *Phishing Detector* is that it is a proprietary plugin. We uploaded all the emails in our study, and the decision changed each time we uploaded the emails again, making it impossible to determine how the label decision is made about emails. Since we want to evaluate the label itself rather than the technical mechanism behind it, we decided to use only the medium variant of the support system, meaning the yellow spam label. Due to a lack of understanding of this support system's inner workings and the fact that all non-deterministic filtering mechanisms are imperfect, we decided to include false positives and negatives. Here, we deviate from the warning banner and external marker and include false negatives for the spam label. For the warning banner and external marker, false negatives were exceptional cases that could be avoided; for the spam label, they are unavoidable. A false negative is a phishing email that is not labeled at all, and a false positive is a legitimate email that is labeled as spam. In 10 % of cases across all participants per email type, we introduce intentional misclassification. We assumed 10 % false-positive and false-negative rates to approximate a somewhat realistic, moderately imperfect system. Technically, we generate a random number X which is sampled uniformly from $[0, 1)$. When $X < 0.10$, we either (i) suppress the spam label for a phishing email (false negative), or (ii) show the spam label for a legitimate email (false positive).

3.2.4 Tooltip. When a user hovers over a link in an email with the mouse, a tooltip appears with the actual URL either at the mouse cursor or at the bottom of the window. Volkamer et al. [55] developed a special tooltip, called TORPEDO, that helps users to read and understand URLs. TORPEDO displays the URL in the tooltip in bold and with increased character spacing and shows an additional risk assessment of the URL with colored borders. A green border indicates low risk for, e. g., popular sites, a blue border signifies previously visited or allowlisted sites, and a gray outline denotes unknown risk. Here, the user must check the URL themselves. A screenshot can be seen in Figure 3b. At the same time, TORPEDO blocks the link for three seconds, and a timer is displayed. The user can only continue and click on the link after the time has passed. The user can adjust the timer manually in the settings. Additionally, one feature of TORPEDO is to resolve links from a link shortener, when the user clicks on the respective button.

In our tool, we only replicated an email client; therefore, installing the TORPEDO plugin[4] was not possible. In addition, participants also could not click on links and visit a website. Therefore, we needed to make a few adjustments to integrate TORPEDO. We assume a user who classified the email as legitimate would have clicked on the link. In that case, we block the action and display the URL, along with a timer set to 3 seconds. After the timer expires, the user can click on the *Legitimate* button. If the user has already hovered over the link before, and the time has expired, the tooltip is not displayed again when they select *Legitimate*. Likewise, if the participant selects *Fraudulent*, the tooltip is not displayed, as we assume that the participant would not then click on the link. Since TORPEDO supports three colored assessments, we only perform this action when the assessment is grey (unknown); in the case of blue and green, the tool does not block that action.

To find out which of our selected test emails gets which colored assessment, we installed TORPEDO in a real email client and tested each email. For an email that leads to the SharePoint page (see Figure 24) of the fictional company, we used the blue version of the tooltip. Since this website does not exist in reality, it would actually have a gray border. The tooltip is displayed for all emails that contain a link. That means we have false positives, because the tooltip is also displayed for legitimate emails, but no false negatives, because the tooltip is shown for all phishing emails, but only in gray.

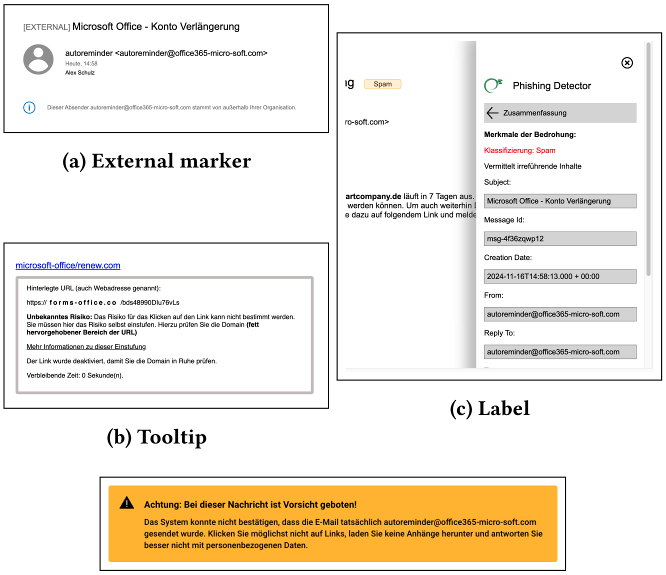
3.3 Email Selection

We conducted a pre-study to select the emails and to determine their difficulty. Table 1 shows an overview of the selected emails, and which support system is displayed in which variation for which email. Additionally, the table shows the level of difficulty, whether links or attachments are included and whether it is phishing or a legitimate email.

3.3.1 Pre-Study. We conducted a preliminary study to select the emails with an early iteration of our interactive tool. In this iteration, we did not include any support systems, and the tutorial did not require participants to open the calendar. We conducted this study with 126 click workers from the German crowd sourcing platform *clickworker.de*, and their task was to classify 26 emails as phishing or legitimate. Out of these, 11 emails represented phishing attacks, and 15 emails were legitimate. When creating these emails, we adapted existing real phishing emails and also used emails from Steves et al. [49] as a basis and adapted them to fit the scenario. We included low-risk emails that have no links or attachments and only share information. Additionally, we ensured that the support systems would be triggered for certain emails, e. g., we included emails from external senders, links to popular or lesser known websites, and attachments. The emails were first rated independently

[3] https://www.emailveritas.com/

[4] https://github.com/SecUSo/torpedo



[EXTERNAL] Microsoft Office - Konto Verlängerung

autoreminder <autoreminder@office365-micro-soft.com>
Heute, 14:58
Alex Schulz

Dieser Absender autoreminder@office365-micro-soft.com stammt von außerhalb Ihrer Organisation.

(a) External marker

microsoft-office/renew.com

Hinterlegte URL (auch Webadresse genannt):

https:// **forms-office.co** /bds48990DIu76vLs

Unbekanntes Risiko: Das Risiko für das Klicken auf den Link kann nicht bestimmt werden. Sie müssen hier das Risiko selbst einstufen. Hierzu prüfen Sie die Domain **(fett hervorgehobener Bereich der URL)**

Mehr Informationen zu dieser Einstufung

Der Link wurde deaktiviert, damit Sie die Domain in Ruhe prüfen.

Verbleibende Zeit: 0 Sekunde(n).

(b) Tooltip

artcompany.de läuft in 7 Tagen aus.
werden können. Um auch weiterhin
e dazu auf folgendem Link und melde

Phishing Detector

Merkmale der Bedrohung:

Klassifizierung: Spam

Vermittelt irreführende Inhalte

Subject: Microsoft Office - Konto Verlängerung

Message Id: msg-4f36zqwp12

Creation Date: 2024-11-16T14:58:13.000 + 00:00

From: autoreminder@office365-micro-soft.com

Reply To: autoreminder@office365-micro-soft.com

(c) Label

Achtung: Bei dieser Nachricht ist Vorsicht geboten!

Das System konnte nicht bestätigen, dass die E-Mail tatsächlich autoreminder@office365-micro-soft.com gesendet wurde. Klicken Sie möglichst nicht auf Links, laden Sie keine Anhänge herunter und antworten Sie besser nicht mit personenbezogenen Daten.

(d) Warning banner

Figure 3: The graphics show an overview of the support systems as we reproduced them, using the example of the email *microsoft*. Subfigure (a) represents the external marker with the label [EXTERNAL] before the subject line and as a note above the email. Subfigure (b) shows the tooltip (with gray border), after hovering over the link. In Subfigure (c) one can see the label and the corresponding opened sidebar, including email header information. Subfigure (d) shows the warning banner which would be displayed above the email content.

by two researchers on the NIST phish scale [49], then discussed and adjusted in terms of their rating again. The 126 participants classified the emails, and their assessment did not match our rating with phish scale [49]. This mainly concerned phishing emails, which we rated as medium to difficult, but which were still rated correctly by the majority of participants. Furthermore, we excluded emails that were too difficult or too easy to detect. Then we selected three phishing emails with the difficulty levels easy (recognized by 75 % or more of participants), medium (recognized by 50 % to 74 %), and difficult (recognized by under 50 %). All legitimate emails were used for the final study, which resulted in 18 emails overall. For these emails, we set the limits for easy emails at 90 % or more of participants recognizing them, for medium emails at 80 % to 89 %, and for difficult emails at less than 80 % recognizing them. We aimed to closely balance the ratio of different emails in the final study to reflect a real-life scenario, with more legitimate emails than phishing emails. The selected emails are described in more detail in the following Sections 3.3.2 and 3.3.3.

3.3.2 Legitimate Email Selection. We selected 15 legitimate emails, including 10 easy, 3 medium and 2 difficult ones, see Table 1. We have classified the difficulty based on the correct classification in the pre-study, i. e., easy ($\geq$ 90 %), medium ($\geq$ 80 %), and difficult ($\leq$ 80 %). Six legitimate emails included links or attachments, and two of them (Doodle and Attachment) are flagged with the gray version of the tooltip, so that the user must check for themselves. The email Resilience contains a link from a popular URL shortener and therefore has a gray tooltip. After clicking on 'Resolve', the URL redirects to a low-risk website and is then transformed into a tooltip with a green border. For the email Password, the tooltip shows a blue border, as it links to an internal website. In two further emails (Dropbox, Feedback), the tooltip has again a green border, as the websites are low-risk. The external marker is displayed for three of the legitimate emails, as these were sent from senders or systems outside the company. The first one is an application, and the other two emails are automatic notifications sent by the external systems Dropbox and Doodle. The warning banner is not displayed on any of the legitimate emails, as we assume that the senders' email servers are configured correctly. Similar to the phishing emails, the label is randomly distributed across all users and displayed as a false positive in 10 % of the emails, which we described in Section 3.2.3.

3.3.3 Phishing Email Selection. We selected three phishing emails with different levels of difficulty (see Table 1). The number of phishing emails is relatively small compared to legitimate emails, with a ratio of 3:15. The reason is that we wanted a more realistic scenario without overwhelming participants with additional emails. According to [27], one-third of emails are unwanted, and 2.3 % contain malicious content. Moreover, [30] reports that approximately 1.2 % of all email traffic is phishing. However, many emails are blocked before they reach the user's inbox, resulting in an even lower ratio of legitimate to phishing emails than in our scenario. While other studies [47, 55] often used a more balanced ratio of phishing and legitimate emails, some approaches are based on more realistic conditions [33, 35, 38, 63]. In addition, we wanted to represent different levels of difficulty, as in reality, many phishing emails are easily detectable. For difficulty levels, we use the NIST phish scale [49] as a basis, which also provides three difficulty levels. We selected the following phishing emails from our pre-study:

Weblogs is a difficult phishing email which we took from [49] and adapted. Only 48.4 % of pre-study participants correctly identified this email. The email alerts users that accessing prohibited websites may result in disciplinary action, so to stay safe, they should verify which websites are flagged by clicking on the provided link. A special case in our scenario is that the email was sent from a compromised internal account, i. e., the sender email address was correct. For this reason, the email is not marked by the external marking or the warning banner. The label marks the email as described in Section 3.2.3. The tooltip classifies the URL as unknown and, therefore, has a gray border. One could correctly classify this email as phishing by hovering over the link.

Microsoft is a medium difficult phishing email which was correctly identified by 72.2 % of participants in the preliminary study. The Microsoft license for the user's account supposedly expires and must be renewed. One sign of phishing is that the URL shown in the content and the actual URL are different. Furthermore, the sender address is also suspicious. The email is marked as external, and the warning banner is shown. Likewise, the tooltip is displayed in the gray version. The label marks the email again as described in Section 3.2.3.

VR-Bank was correctly classified as phishing by 90.5 % of the participants in the preliminary study. Therefore, it poses as an easily detectable phishing email. The email claims to come from a German banking institution and reports increased unauthorized access to accounts. For this reason, users should check the security settings for their accounts. In this case, the external marker, the warning banner, and the gray tooltip are displayed. Again, the label is displayed as described in Section 3.2.3.

3.4 Main Study

Similar to the preliminary study, we recruited participants via the German crowd-sourcing platform clickworker.de for the main study. We excluded all clickworkers who had already participated in the preliminary study. Another requirement was that clickworkers had to be German-speaking and could only participate using a desktop computer or laptop, as the study was not designed for mobile devices. We implemented technical measures to prevent clickworkers with insufficient screen resolution from participating in our study. When detected, a message was displayed stating that participation was not possible and advising the participant to use a different device with a higher screen resolution.

The main study started with a welcome screen that explains the study and the scenario. When the participant clicked "Start" they saw the tutorial as described in Section 3.1.2 and the tool assigned the participant to one of the five test groups, consisting of four support systems (warning banner, external marker, label, and tooltip) and a control group, without a support system. We assigned the support systems to the test groups using a round-robin procedure to ensure an equal distribution. The participant groups with the support system were not informed of it, as we wanted to avoid expectation effects and examine whether a support system really helps, not whether the participants consciously pay attention to it. After the tutorial, the actual experimental part of the study began. The participants saw the 18 emails selected as described in Section 3.3, one after the other in random order. For each of the emails, participants should decide whether it is *legitimate* or *fraudulent* and indicate this by clicking the appropriate button. Depending on the test group, participants were presented with the appropriate support system in the matching emails, as shown in Table 1. Participants could hover over links or open the calendar, contact list, and job offer document to receive more information. In the background, we track all these activities of each participant for the subsequent evaluation. After the participants classified all emails, they see the correct solution together with their results.

3.5 Survey

Once the participants had classified all 18 emails, the tool presented the concluding survey, see Section D. At the beginning of the survey, we asked participants which elements of emails they paid the most attention to in the survey. Secondly, we asked what elements of emails they generally pay attention to when they receive suspicious emails. Both questions were asked as priority rankings, where answers are provided but participants can also add their own answers. Not all items had to be selected. As a basis for these two questions, we used Hasegawa et al. [24]. For the first question, we also listed the support systems for the corresponding groups. If the participant belonged to a group with a support system, we asked whether the participant had noticed the support system at all. If the answer to that question was yes, or at least partially, we asked them to describe their first impressions in a free-text field. Afterwards, we asked the participants to rate the support system on a five-point Likert scale (from "Strongly disagree" to "Strongly agree") how helpful the support system was, whether it had an impact on their decisions, whether they trusted it, and whether they would use it professionally and privately. The next part of the survey was the UEQ-S [45] (a short version of UEQ [44]), a standardized instrument designed to evaluate the user experience. It captures both pragmatic aspects, such as efficiency and perspicuity, as well as hedonic aspects such as stimulation and novelty. Thereby, the participants rated eight pairs of contradictory adjectives (e. g., 'complicated' versus 'easy') on a seven-point Likert scale. Finally, the participants should rate their general impression of the support system on a 10-point scale ("I do not like it at all" to "I really like it") and justify their rating in a free text field. The last page of the survey was again visible to all participants. It contained general demographic questions as well as questions about their affinity for technology and which email clients they use. Additionally, there was an option to enter final comments in a free-text field.

3.6 Data Analysis

In this section, we describe the statistical tests we have used to analyze the quantitative data and the procedure to analyze the qualitative data.

Quantitative Data. The quantitative data did not follow a normal distribution. For this reason, we performed non-parametric Kruskal-Wallis tests to examine differences between the individual support system groups and the control group. If the Kruskal-Wallis test indicates significant differences between groups, we additionally perform pairwise post-hoc comparisons using the Mann-Whitney U (MWU) test to determine between which groups these differences exist. To account for multiple hypothesis testing and to control the family-wise error rate, we applied the Holm–Bonferroni step-down correction to the p-values. We considered p-values below 0.05 to be statistically significant. We used r as the effect size and interpret it as follows: $r < 0.3$ as small, $0.3 \leq r < 0.5$ as medium, and $r \geq 0.5$ as large. In the case of binary variables, we used Fisher's exact tests, also with a Holm-Bonferroni correction and a significance level of $\alpha = 0.05$. The effect sizes were derived from the odds ratios (OR) and categorized as follows: OR = 1 – no effect, ORs between 1 and 1.5 – small, ORs between 1.5 and 2.5 – medium, ORs between 2.5 and 4.0 – large, and ORs ≥ 4.0 – very large. To analyze correlations between demographic data and experimental results, we used Spearman's rank correlation. Furthermore, we used Wilcoxon signed-rank tests to analyze the effects of false positives or false negatives on the classifications of the participants. Finally, we created a logistic regression model to find out which variables had the most significant influence and effect on participants correctly classifying an email.

Table 1: The table shows the rating regarding the degree of difficulty of the emails from the pre-study and the visibility for each support system. The color of the tooltip indicates which variant is displayed. The control group did not have any support system. The P before the email designation stands for phishing and the L for a legitimate email. The ● icon indicates that the support system is shown, ○ indicates that a support is not shown, and ◐ indicates that a support system is only partially displayed; the percent indicates the percentage of cases it is displayed. The details regarding the partially displayed label can be found in Section 3.2.3.

Email	External	Banner	Label	Tooltip	Rating	Link/Attachment
P Weblogs	○	○	◐ (90 %)	● (Grey)	Difficult	Link
P Microsoft	●	●	◐ (90 %)	● (Grey)	Medium	Link
P VR Bank	●	●	◐ (90 %)	● (Grey)	Easy	Link
L Password	○	○	◐ (10 %)	● (Blue)	Difficult	Link
L Doodle	●	○	◐ (10 %)	● (Grey)	Difficult	Link
L Dropbox	●	○	◐ (10 %)	● (Green)	Medium	Link
L Attachment	○	○	◐ (10 %)	● (Grey)	Medium	Link/Attachment
L Resilience	○	○	◐ (10 %)	● (Grey/Green)	Medium	Link/Attachment
L Call Back	○	○	◐ (10 %)	○	Easy	Attachment
L Feedback	○	○	◐ (10 %)	● (Green)	Easy	Link
L Job Application	●	○	◐ (10 %)	○	Easy	Attachment
L Office	○	○	◐ (10 %)	○	Easy	Attachment
L Construction	○	○	◐ (10 %)	○	Easy	-
L Heating	○	○	◐ (10 %)	○	Easy	-
L Present	○	○	◐ (10 %)	○	Easy	-
L Report	○	○	◐ (10 %)	○	Easy	Attachment
L Cake	○	○	◐ (10 %)	○	Easy	-
L IT-Service	○	○	◐ (10 %)	○	Easy	-

Qualitative Data. Two researchers analyzed the answers to the open-ended questions, first separately and then jointly. Both researchers coded the answers independently, and then they met together to create a shared codebook. Based on the shared codebook, the two researchers independently recoded all answers, and then the intercoder reliability was determined using the Brennan–Prediger coefficient [10]. The first open-ended question concerning the reaction yielded $\kappa_{BP} = 0.52$ with a 90 % code overlap threshold, representing a moderate level of agreement. For the second open-ended question regarding the rating, we obtained a Brennan–Prediger coefficient of $\kappa_{BP} = 0.50$ with 90 % code overlap. We then revised the codebook again by combining similar codes. Lastly, we used a consensus-based approach, where both researchers discussed each coding together and agreed on a coding.

3.7 Ethics

Participants were recruited for both, preliminary and main studies, via the German crowdworking platform *clickworker.de* which adheres to GDPR. Our institutions do not have ethics review boards for non-medical studies, therefore, we addressed ethical concerns as follows. Registration on the platform requires participants to be at least 18 years old. Participants were informed that the data collection is for scientific purposes. We only received anonymized data linked to an ID, while the clickworker platform ensures compliance with all remaining data protection requirements. For further questions the participants were provided with an email address. They could exit the study at any time without adverse effects. The study was hosted on a server belonging to the first author’s organization. Answering demographic questions was voluntary. The collected data was stored on a secure server, where only project members had access. Participants received compensation of EUR 3.20, for an estimated study duration of 10 min to 15 min, which corresponds to the German minimum wage of EUR 12.82 per hour from January 2025.

3.8 Participants

The number of participants is based on an a priori power analysis with one-way ANOVA. We set the significance level at $\alpha = 0.05$, the desired power at 0.8 and used Cohen’s f as the effect size. The analysis showed that to detect a large effect $f = 0.4$, approximately 16 participants per group are required, for a medium effect $f = 0.25$, 40 participants per group, and for a small effect $f = 0.1$, 240 participants per group. Consequently, we decided on 100 participants per group (total of 500 participants) in order to ensure adequate statistical power after incomplete or invalid data had been filtered out. Out of a total of 500 clickworkers, the results of 453 clickworkers were included in the evaluation. Participants were excluded if the same IP address appeared multiple times, if incorrect or missing values appeared in the evaluation, if they repeatedly selected the same answer option, or if they completed the study unusually fast. For this purpose, we excluded all participants whose completion time was below the 2.5th percentile. All participants above the 97.5th percentile were retained. Overall, the average time needed by participants was 12.6 min.

The distribution of support systems among participants is as follows: Banner - 87, External - 99, Label - 85, Tooltip - 91, Control

Group - 91. Slightly more than one-third (36.9 %) of participants are female and nearly two-thirds are male (62.5 %). Three participants did not specify their gender. The majority of participants belong to the 20 to 49-year-old age groups. More than half of the participants (56.1 %) are working as employees, followed by 21.4 % who are self-employed and 9.3 % who are university students. Participants without a computer science background make up the majority with 69.8 %, in contrast to participants with a background in computer science (28.5 %). According to a self-assessment, 57.4 % are moderately IT-affine, 35.1 % are highly IT-affine, 6.6 % are not very IT-affine, and 0.9 % did not respond. More details on demographics can be found in Table 5 (Section B).

4 Results

In the following sections, we first present the experimental results, which answer RQ1-3 in Section 4.1, and then answer RQ4 in Section 4.3 by presenting survey results.

4.1 Experimental Results

4.1.1 General Influence of Support Systems (RQ1). Out of the 18 emails, participants correctly identified, on average, between 88.3 % and 83.9 %, depending on the group (see Figure 4). A Kruskal-Wallis test revealed significant differences between the groups ($H(4) = 13.5$, $p < 0.01$) with a small effect size ($\varepsilon^2 = 0.02$). Subsequent Bonferroni-Holm-corrected MWU tests among the individual groups showed significant differences between the warning banner group and the tooltip group ($U = 4\,913$, $p_{\text{Holm}} < 0.05$), with a small effect size ($r = 0.21$). Considering only medium- and difficult-level emails, only the label and tooltip showed significant differences, with small effect sizes. The corresponding MWU results can be found in Section C.1.

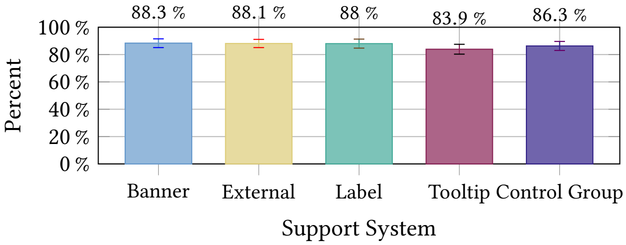


Figure 4: Results of the different test groups for the classification of fraudulent or legitimate emails. The error bars show 95 % confidence intervals.

When dividing the emails into phishing and legitimate, a Kruskal-Wallis test revealed a marginally significant difference with a small effect size between the support system groups for legitimate emails ($H(4) = 9.7$, $p < 0.05$, $\varepsilon^2 = 0.013$). However, after the Bonferroni-Holm correction, we found no significant differences in the post-hoc MWU tests. In the case where only phishing emails were considered, no significant difference was found in the Kruskal-Wallis test ($H(4) = 8.6$, $p \geq 0.05$, $\varepsilon^2 = 0.010$). This shows that support systems did not make a practical impact on email classification in our scenario.

4.1.2 Overcautious Behavior (RQ2).

Individual Emails. To investigate overcautious behavior, i. e., whether using support tools leads to a greater proportion of legitimate emails being classified as phishing than in the control group, we first examine individual legitimate emails. Table 2 shows the number of participants who correctly classified an email across the different support system groups. We performed Fisher's exact tests for each email and support system group with p-values corrected for multiple testing.

The legitimate email about password change was challenging to classify correctly across all groups, with no significant differences between them. Participants in the tooltip group performed worse at classifying legitimate email Attachment and Doodle. The difference was statistically significant for the Attachment email for all groups, and specifically for the control group with a large effect size: $p_{\text{Holm}} < 0.05$, $OR_{\text{control vs. tooltip}} = 0.38$, 95 % CI = $[0.20, 0.72]$. For the Doodle email, there was a statistically significant difference only between external versus tooltip groups: $p_{\text{Holm}} < 0.01$, $OR_{\text{external vs. tooltip}} = 0.34$, 95 % CI = $[0.18, 0.63]$. In both emails, the tooltip had a gray border, indicating an unclear risk assessment, suggesting that participants had difficulty assessing the risk, even with the tooltip's assistance. We did not find statistically significant differences for all other emails and groups.

Displayed Support Systems for Legitimate Emails. To further investigate overcautious behavior, we examine the classification results for legitimate emails that displayed support systems. An external marker was displayed deterministically for 3 legitimate emails from an external server. The warning banner was shown deterministically only for phishing emails. The tooltip was shown for all emails with links, including 6 legitimate emails, and it provided an unclear risk assessment for three of them. Spam label, a probabilistic mechanism, exhibited a more complex behavior: It was flagged as a false positive in 10 % of all legitimate emails. The distribution of false positives was as follows: 0× among 23 participants, 1× among 25 participants, 2× among 19 participants, 3× among 14 participants, 4× among 3 participants, and 5× for 1 participant.

There were no significant differences between the group with the external marker and the control group with the same emails. This also applies to the group with the tooltip and the same emails in the control group. For the label, it is not possible to separate individual emails, as they were displayed across all emails, as described above. Therefore, we can only compare the label group and emails where it was displayed with the entire control group, as each email in the label group was tagged multiple times with a label. However, as the label group has 1.5 false positives on average and the control group has 15 legitimate emails, statistical comparison does not make much sense here due to the large difference in the number of classified emails between the groups. Additionally, we already compared the classification of legitimate emails across all participants in the label and control groups in Section 4.1.1 and found no statistically significant difference.

We conclude that we found evidence of overcautious behavior only in the tooltip group for individual emails.

4.1.3 Influence of False Positives and False Negatives in Support Systems (RQ3). In our scenario, genuine false positives only exist

Table 2: The table shows how often an email was correctly classified by each support system group, i. e., legitimate emails as legitimate and phishing emails as fraudulent. E/M/D means Easy/Medium/Difficult; L means link, and A means attachment as part of the email.

	Email	Banner	External	Label	Tooltip	Control Group
Legitimate	Attachment (M;L+A)	85.1 %	82.8 %	83.5 %	57.1 %	78.0 %
	Report (E;A)	96.6 %	97.0 %	91.8 %	96.7 %	93.4 %
	Job Application (E;A)	93.1 %	91.9 %	91.8 %	85.7 %	90.1 %
	Doodle (D;L)	70.1 %	75.8 %	67.1 %	51.6 %	69.2 %
	Dropbox (M;L)	88.5 %	84.8 %	92.9 %	86.8 %	91.2 %
	Feedback (E;L)	87.4 %	92.9 %	91.8 %	85.7 %	84.6 %
	Present (E)	98.9 %	100.0 %	98.8 %	98.9 %	98.9 %
	Heating (E)	97.7 %	96.0 %	95.3 %	95.6 %	97.8 %
	Cake (E)	98.9 %	99.0 %	97.6 %	93.4 %	96.7 %
	Password (D;L)	50.6 %	54.5 %	60.0 %	57.1 %	45.1 %
	Resilience (M;L+A)	82.8 %	84.8 %	77.6 %	75.8 %	85.7 %
	Call Back (E;A)	92.0 %	86.9 %	87.1 %	82.4 %	86.8 %
	IT Service (E)	95.4 %	98.0 %	90.6 %	94.5 %	94.5 %
	Construction (E)	98.9 %	98.0 %	96.5 %	97.8 %	100.0 %
	Office (E;A)	96.6 %	94.9 %	89.4 %	91.2 %	91.2 %
Phishing	Microsoft (M;L)	89.7 %	82.8 %	91.8 %	86.8 %	78.0 %
	VR (E;L)	93.1 %	88.9 %	96.5 %	92.3 %	91.2 %
	Weblogs (D;L)	74.7 %	76.8 %	83.5 %	80.2 %	80.2 %

for the label and the tooltip as described in Section 4.1.2. All other support systems have no false positives. If we compare the conditions for the label group before the first false positive was displayed and the conditions afterwards (excluding the first false positive email), there is no significant difference in the correct classification of the emails, as shown by a Wilcoxon pre-post test (before: 87.8 %, after: 87.9 %; $W = 431.5$, $p \geq 0.05$, $r = 0.03$). For the tooltip, there is a slight increase in correctly classified emails after the first false positive email, but it is not statistically significant (before: 77.7 %, after: 84.1 %; $W = 1\,387.5$, $p \geq 0.05$, $r = 0.08$). In our scenario, the false positives had no noticeable effect on participants' subsequent decisions.

False negatives occurred for three support systems: deterministically via a warning banner and an external marker for the Weblogs phishing email, and probabilistically in 10 % of phishing emails for the label, resulting in 28 participants not seeing a label in one phishing email. Fisher's exact test did not reveal statistically significant differences for the Weblogs email. However, the percentages of correct classification are slightly lower for banner and external (see Table 2). This means there is a slight tendency to trust the support system too much when indicating suspicious emails. For the label and warning banner support systems, there was no significant change in participants' email classification after the first email was marked as a false negative, as shown by a Wilcoxon pre-post test. For the external marker, the correct classification by the participants increased slightly after they saw the false negative email (before: 85.1 %, after: 90.5 %; $W = 888.5$, $p < 0.05$, $r = 0.22$). Thus, false negatives also had little impact in our study.

We performed robustness checks to assess the impact of randomly introduced misclassifications for the label group, where 10 % of all emails across all participants were misclassified. This would mean an average of 1.5 false positive emails and 0.3 false negative emails per participant, with 85 participants in this group. For robustness checks, we considered only participants with the closest number of classification errors (0 or 1 false negatives and 1 or 2 false positives). Thus, we carried out MWU tests using only the results of the previously described label group participants and compared them with all other groups, regarding phishing or legitimate emails. We found only minor differences, which can be attributed to differences in the number of participants in this group (see Section C.1).

4.2 Influence of Additional Aspects

In this section, we analyze whether specific aspects influence the classification results. In practice, decisions about potentially threatening emails are not made in isolation, but always within a situational context. For this reason, we have developed and integrated a corresponding scenario. At the same time, we are investigating whether the participants' demographic characteristics influenced the results.

4.2.1 Contextual Factors. To determine whether contextual factors influence the correct classification of an email, a logistic regression was performed; see Figure 5. When all independent variables are set to their reference levels and continuous variables are set to zero, we estimated the probability of correctly classifying an email to be 88.5 %. We use emails of medium difficulty as a reference.

For difficult emails, the probability decreased significantly to 75.4 % ($\beta = -0.92$, OR = 0.40, 95 % CI [0.34, 0.47], $p < 0.001$). For easy emails, it significantly increased to 94.0 % ($\beta = 0.72$, OR = 2.05, 95 % CI [1.59, 2.65], $p < 0.001$).

There was a significant difference when a participant opened the contact list ($\beta = 0.58$, OR = 1.79, 95 % CI [1.49, 2.15], $p < 0.001$), with the probability of correctly classifying an email increasing to 93.2 % compared to not opening the contact list. If a participant opened the calendar, it slightly increased the probability for a correct classification (89.0 %) in reference to not opening the calendar, but the effect was not statistically significant ($\beta = 0.06$, OR = 1.06, 95 % CI [0.82, 1.36], $p \geq 0.05$). For participants who opened the job offer the probability of a correct classification decreased to 84.0 % ($\beta = -0.38$, OR = 0.69, 95 % CI [0.50, 0.95], $p < 0.05$) in comparison to those who did not open it. The job offer is the least frequently used resource of all.

We further observed that hovering over a link significantly increased the probability of an email being correctly classified to 91.2 % compared to not hovering ($\beta = 0.30$, OR = 1.35, 95 % CI [1.14, 1.59], $p < 0.001$). The presence of a link in an email leads to a decreased probability to 79.9 % ($\beta = -0.658$, OR = 0.52, 95 % CI [0.40, 0.68], $p < 0.001$) compared to emails without links.

Concerning support systems, there is a slight decline in correct classifications to 86.1 % in the tooltip group ($\beta = -0.22$, OR = 0.81, 95 % CI [0.66, 0.99], $p < 0.05$) compared to the control group. For the other support systems, we found an increase from 89.9 % to 90.4 %, but it was not statistically significant. This corroborates the results of Section 4.1.

We found that as processing time increases, the probability of correctly classifying an email decreases, with a minimal effect. A 10-second increase in the processing time for an email results in a decrease to 88.1 % ($\beta = -0.04$, OR = 0.96, 95 % CI [0.94, 0.98], $p < 0.001$).

Support systems played a minor role in determining whether an email was correctly or incorrectly classified. Instead, contextual factors, especially the contact list, had a greater influence, as did the presence of a link and whether the participants hovered over it.

To ensure the robustness of the results, we conducted robustness tests by re-estimating the model after removing the contextual variables, which did not yield significant changes. In addition, we refitted the model with heteroskedasticity-consistent standard errors (HC3) and found only minor differences.

4.2.2 Correlations with Demographics. For this analysis, we excluded participants who did not respond to a specific demographic question. We found a significant difference between male and female participants in the number of correctly classified emails ($U = 26410.0$, $p < 0.05$), although the effect size is small ($r = 0.10$). Among male participants, the average number of correctly classified emails was 15.8, and among female participants, it was 15.4. The correlation between participants' age and the number of correctly classified emails was weakly negative. It did not reach statistical significance ($\rho = -0.09$, $p \geq 0.05$), as analyzed using a Spearman rank-order correlation. However, there were significant differences between the amount of correctly classified emails and the participants' own assessment of their IT affinity ($H(2) = 15.5$, $p < 0.001$, $\varepsilon^2 = 0.03$). An MWU test showed significant differences with a small effect between the strongly affine and average affine groups ($U = 24\,583.0$, $p_{\text{Holm}} < 0.01$, $r = 0.16$) and between the strongly affine and little affine groups ($U = 3\,187.0$, $p_{\text{Holm}} < 0.01$, $r = 0.23$). Participants who considered themselves as having little IT affinity correctly classified an average of 14.9 emails. Average IT-affine participants correctly classified an average of 15.5 emails, while those with strong IT affinity classified an average of 16.1 emails correctly. Overall, demographic factors had little influence on classification results.

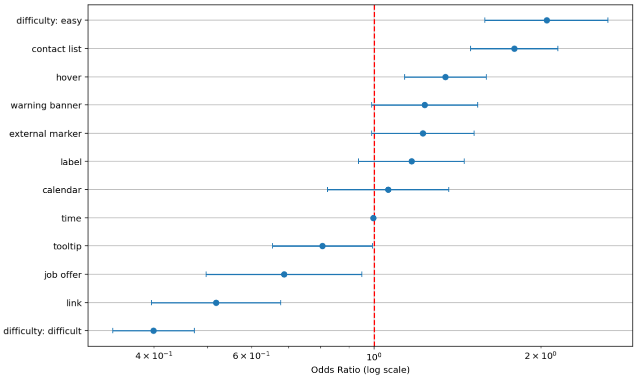


Figure 5: Visualization of the estimated odds ratios with 95 % confidence intervals in a forest plot (log scale), with a vertical reference line at OR = 1 indicating no effect compared to the reference level.

4.3 Perception of Support Systems by Users (RQ4)

This section analyzes the results of our survey. We refer to the participants with a unique identifier consisting of *P* and a sequential number.

4.3.1 Priority of Phishing Cues. We first asked which part of the email the participants paid most attention to, both in the study and in general. Figure 28 in Section C.2.1 shows the results of the two priority rankings per support system (study/general) in detail.

Sender's email address is mentioned most frequently across all participants, with 415 mentions (91.6 %). *URLs in emails* (396 - 87.4 %) and the *Sender's name* (360 - 79.5 %) follow. A total of 72 participants out of 87 participants selected the warning banner in the corresponding group (82.8 %).

The least frequently selected were *Own involvement* with 239 times (52.8 %) and *Urgency of email* (241 - 53.2 %). Also in this range are the support systems external marker, spam label and tooltip, which were selected by 56.5 % to 62.6 % of participants within their groups.

Furthermore, six participants contributed their own elements, for example, the "feeling" (P 138) when looking at the email or the "time of the email" (P 50).

4.3.2 How Many Noticed the Support System. Subsequently, we asked all participants in the support system groups whether they had noticed the support system (see Figure 6). The warning banner was the most frequently noticed support system at 94.3 %. The tooltip was the second-most-frequently noticed support system, with 75.8 % noticing it. The label and the external marker were noticed by 60.0 % and 57.6 %, respectively. We conclude that most participants who noticed their respective support system also utilized it in their decisions, but not all of them (cf. Section 4.3.1).

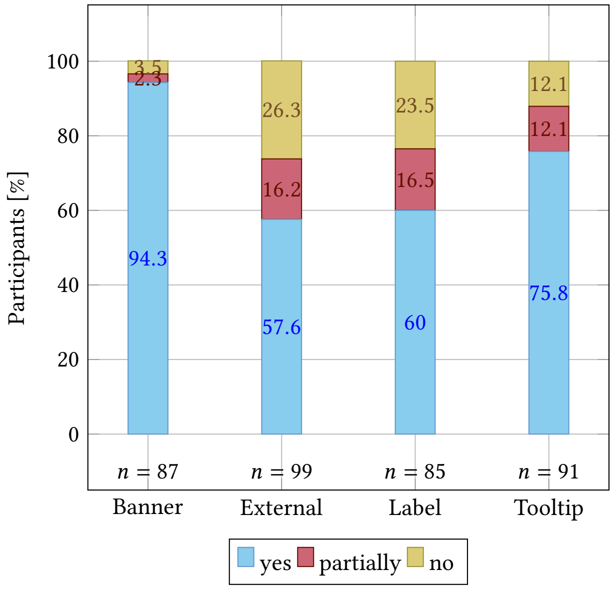


Figure 6: How many participants in the different support systems groups noticed the support system during the study.

The following questions, until the demographic section of the survey, were only asked to participants who had at least partially noticed the support system, which reduces the number of participants per support system to the following: banner with 84 participants, external marker with 73 participants, label with 65 participants, and tooltip with 80 participants.

4.3.3 Initial Reaction Seeing the Support System. We did not inform participants about the support system before the experimental part to avoid expectation effects and to determine whether the system was understood intuitively. We asked each participant to describe their initial reaction upon seeing the support system in an open-ended question to identify their spontaneous impressions. In Table 3, we show the distribution of codes as a percentage of the total number of participants assigned to the respective support system. Overall, we defined 16 codes, distributed across the categories *positive* (green), *neutral* (blue), *negative* (red), and *other* (grey).

Most popular positive statements for external marker, banner, and label were that the support system *increased awareness* and encouraged participants to further *check the email*, or that it caused *mistrust* towards the email and provided *hints*. For example, P54 stated that the warning banner meant to them was that they "definitely needed to read [the email] more carefully". The tooltip stands out as the most popular positive statement for it: it is *helpful/useful*, but without explaining why (e. g., P342: "Helpful add-on.").

Between 6.8 % and 10.7 % of participants mentioned that they *blindly trusted* the support system. For example, P371 stated that they automatically assumed that this was phishing when they saw the warning banner. The tooltip was especially often found *unclear*.

Among the negative codes, participants often reported feeling *uncertain/puzzled/confused* about the support system. For example, P210 said, referring to the external marker: "At first, I was surprised and concerned about the potential implications of that." Compared to all support systems, participants most frequently criticized the label's accuracy and reliability. For example, P200 said: "I ignored it. In general, too often incorrect." This could be explained by the presence of false positives in this case. The tooltip was more often found "not (very) helpful" than other support systems. This might be because the tooltip repeatedly gave an uncertain risk assessment, leaving the responsibility for checking to the participant, as P56 described: "Not very helpful, as the risk was mostly unknown except for already known senders."

4.3.4 Characteristics of the Support System. In the next section of the survey, participants rated the support system's characteristics on a 5-point Likert scale and noted whether they would use it for professional or personal purposes. Figure 7 shows the results as box plots for each support system.

Helpfulness was rated the highest for the warning banner (median 5), which is the deterministic support system without false positives, although it exhibited one false negative. For all other support systems, the median was 4, which is also relatively high, given their respective weaknesses, which were qualitatively criticized (see Sections 4.3.3 and 4.3.6).

With respect to the impact of support systems on decisions, the medians for the groups' warning banner, external marker, and tooltip are at 4, and the label has a median rating of 3. Likewise, the median rating for the trustworthiness was 4 for the banner, external marker, and tooltip groups, and 3 for the label group. Thus, the spam label as a support system performed the worst, likely due to its probabilistic nature.

The statement as to whether participants would use the support system in a work-related context was rated with a median of 4 in each group. Whether participants would use it privately also received a median rating of 4 in the groups, with only the external marker receiving a median rating of 3, which is understandable, as it is only useful in the context of organizations.

4.3.5 UEQ-S Results. The UEQ-S can be interpreted as follows: Values between -0.8 and 0.8 represent a neural evaluation of the corresponding scale, values > 0,8 represent a positive evaluation, and values < -0,8 represent a negative evaluation [43]. Results reveal that all support systems received positive pragmatic quality scores, with the banner ranking highest (mean = 1.7, SD = 1.3), followed by the external marker (1.6, 1.2), label (1.3, 1.1), and tooltip (1.0, 1.5). Hedonic quality scores were neutral across all systems, with the tooltip leading the way. The external marker achieved the highest overall score, followed closely by the warning banner, tooltip, and label, indicating slightly positive evaluations for the external marker and warning banner, and neutral ratings for the label and tooltip. The external marker achieved the highest overall score (1.0, 0.9), followed closely by the warning banner (0.9, 1.0), tooltip (0.8, 1.2), and label (0.7, 0.9), indicating slightly positive evaluations for the external marker and warning banner, and neutral ratings for the label and tooltip. Figure 8 presents the UEQ benchmark chart based

Table 3: Distribution of codes for participants' initial reactions as a percentage of the total number of participants assigned to each support system. Numbers in brackets denote the number of responses. Some responses were assigned multiple codes.

Codes	External (73)	Banner (84)	Label (65)	Tooltip (80)	
increased awareness / checking email	38.4 %	39.3 %	24.6 %	17.5 %	positive
mistrust of email / hint	20.5 %	26.2 %	33.8 %	8.8 %	
helpful / useful	5.5 %	3.6 %	3.1 %	31.3 %	
identified as suspicious	1.4 %	10.7 %	4.6 %	2.5 %	
delete / ignore email	0.0 %	3.6 %	3.1 %	0.0 %	
confirmation of own estimation	0.0 %	2.4 %	1.5 %	1.3 %	
blindly trusted / unreflective	6.8 %	10.7 %	9.2 %	10.0 %	neutral
neutral / unclear	4.1 %	6.0 %	3.1 %	15.0 %	
conventional	4.1 %	1.2 %	0.0 %	1.3 %	
uncertain / puzzled / confused	8.2 %	3.6 %	3.1 %	8.8 %	negative
reliability / accuracy	0.0 %	0.0 %	20.0 %	2.5 %	
ignored	5.5 %	1.2 %	3.1 %	2.5 %	
mistrust of support system	1.4 %	6.0 %	3.1 %	1.3 %	
not (very) helpful	0.0 %	1.2 %	1.5 %	7.5 %	
danger	0.0 %	1.2 %	1.5 %	2.5 %	
(visual) design	0.0 %	0.0 %	0.0 %	3.8 %	
unclear	5.5 %	2.4 %	4.6 %	1.3 %	Other

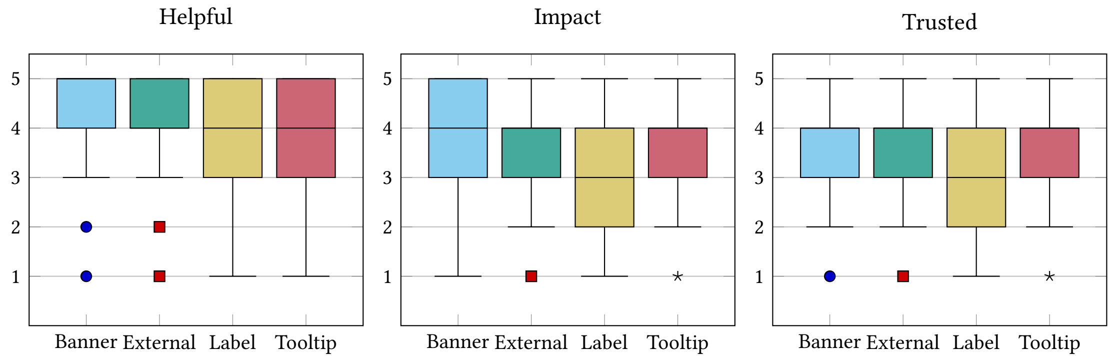
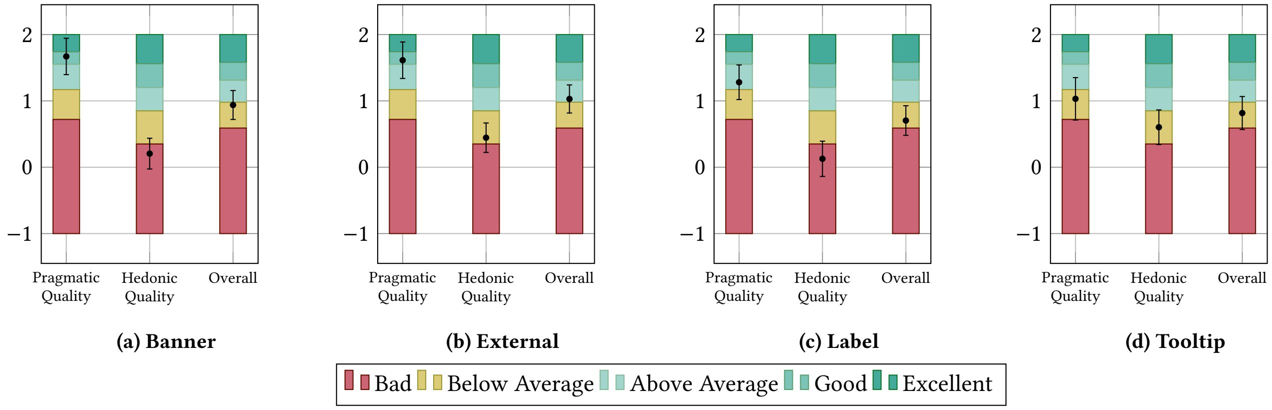


Figure 7: The graphs show, as box plots for each support system, how helpful the support system was, what impact the support system had, and the extent to which participants trusted the support system.

on 452 product evaluations from the official UEQ website [44]. Note that this data derives from the full UEQ rather than the short version used in our study, warranting cautious interpretation. The tooltip scored *below average* across all scales. For pragmatic quality, the warning banner and external marker were rated as *good*, while the label was *above average.* Hedonic quality ratings were less favorable, with the tooltip and external marker scoring *below average*, and the warning banner and label falling into the *bad* range. Overall, only the external marker approached the *above average* threshold, while all other systems remained *below average.*

4.3.6 Overall Impression. Participants rated their overall impression of the support system on a 10-point Likert scale from "I don't like it at all" to "I really like it" and then justified their opinion in an open-ended question. The banner and the external marker are equally rated with a median of 8. Both the tooltip and the label have a median rating of 7, see also Figure 9. Overall, the general impression results mirror the results of the previous Section 4.3.4, where the warning banner and external marker slightly outperform the spam label and the tooltip.

Participants were also asked to justify their opinions in an open-ended question. Table 4 shows the distribution of codes as a percentage of participants assigned the corresponding code per support system. There are a total of 22 codes, which are divided into the categories *positive* (green), *neutral* (blue), *negative* (red), and *other* (grey).



(a) Banner

(b) External

(c) Label

(d) Tooltip

Figure 8: UEQ-S results on benchmark charts for our investigated support systems, namely banner (a), external marker (b), label (c), and tooltip (d) with the mean values for Pragmatic Quality, Hedonic Quality, and the Overall Result. The error bars show the confidence intervals (p = 0.05) per scale.

Table 4: The table the shows distribution of codes as a percentage based on the total number of participants assigned per support system regarding the reasons for for the participants' rating. Numbers in brackets denote the total number of responses. Many responses were coded with more than one code.

Codes	External (73)	Banner (84)	Label (65)	Tooltip (80)	
helpful/useful	48.0 %	22.6 %	35.4 %	55.0 %	positive
increased awareness/felt alerted	17.8 %	27.4 %	15.4 %	8.8 %	
efficiency	12.3 %	6.0 %	0.0 %	6.2 %	
likable	8.2 %	11.9 %	7.7 %	7.5 %	
conspicuous	6.9 %	28.6 %	9.2 %	0.0 %	
simple/understandable	4.1 %	8.3 %	3.1 %	6.2 %	
skeptical / uncertain	12.3 %	8.3 %	18.5 %	10.0 %	neutral
conventional	8.2 %	8.3 %	3.1 %	2.5 %	
unclear/neutral opinion	5.5 %	3.6 %	6.2 %	2.5 %	
sense of security	4.1 %	4.8 %	0.0 %	2.5 %	
unfamiliar/new	1.4 %	0.0 %	0.0 %	5.0 %	
improvements	4.1 %	7.1 %	13.8 %	10.0 %	
inconspicuous	8.2 %	1.2 %	1.5 %	1.2 %	negative
unnecessary	5.5 %	3.6 %	6.2 %	5.0 %	
reliability/accuracy	2.7 %	2.4 %	16.9 %	12.5 %	
comprehensibility	2.7 %	2.4 %	1.5 %	2.5 %	
danger	2.7 %	2.4 %	1.5 %	5.0 %	
(visual) design	1.4 %	1.2 %	0.0 %	5.0 %	
disruptive	1.4 %	2.4 %	0.0 %	2.5 %	
frightening	1.4 %	3.6 %	3.1 %	0.0 %	
complicated/cumbersome	0.0 %	0.0 %	1.5 %	3.8 %	
Unclear	4.1 %	2.4 %	1.5 %	0.0 %	Other

All systems are described as *helpful/useful* by many participants, with the warning banner also described as *conspicuous* most often (e. g., P417: "Stands out instantly and immediately catches the eye."). Warning banner is especially often complimented for *alerting* its users to pay more attention to the respective email, and external marker for increasing *efficiency* in email processing (e. g., P446: "Speeds up finding emails within the company and sorting out external ones at a glance.").

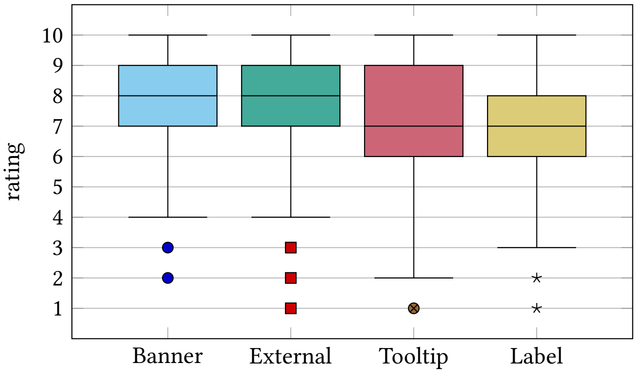


Figure 9: The figure shows box plots of the participants' ratings for each of the four support systems.

Participants most often commented that they were "skeptical/uncertain" about the support system for systems with false positives (label) or with unclear assessments that need further investigation (external marker, tooltip). For example, P130 about external marker: "It's good, but you shouldn't blindly trust it." Some participants suggested *improvements*, e. g., P164 about label: "Generally helpful, but more explanation as to why something was marked as spam would be even more helpful."

Overall, we found the fewest negative statements concerning the banner. Label and tooltip received the most negative comments regarding their "reliability/accuracy". P365 describes their criticism of the tooltip as follows: "Too much text, too many values, and inaccuracies. In the end, I decide for myself based on my experience, overall impression, common sense, and gut feeling." Furthermore, participants criticized the *(visual) design*, considered the tooltip *unnecessary*, or feared it could be a *danger*. P58 implied that the assessment of the tooltip concerns only the URL and the actual phishing attempt that could be hiding behind it. Therefore, the tooltip could induce a false sense of security in people: "The risk of falling into a trap specifically because of the tooltip is relatively high. As soon as someone uses supposedly secure third-party providers (Google, Dropbox, etc.) in the context of phishing, the tooltip suggests that it is safe per se." External marker was most criticized for being hard to notice, and for habituation effect, e. g., P49: "I constantly receive emails from 'external parties', so I would quickly stop noticing this marking."

To summarize, systems with false positives or unclear assessments that require further thought are perceived as more negative and less helpful than deterministic systems without false positives. False negatives have a lesser effect on the rating.

4.3.7 Correlations of Experimental Results with the Overall Impression Rating. We conducted a Spearman rank-order correlation to investigate the relationship between the rating of the support system and participants' experimental results. The analysis revealed a weak, negative correlation that was statistically significant ($\rho = -0.146$, $p < 0.05$). This means that participants with fewer correctly classified emails tended to give a slightly higher rating. When considering the support system groups individually, significant correlations were found for the label ($\rho = -0.349$, $p < 0.01$, moderately negative) and external marker ($\rho = -0.279$, $p < 0.05$, weakly negative), which also shows that higher ratings are linked to less correct results. There were no significant correlations between tooltips or banners and the number of correct results in the experimental part. Thus, there is evidence that a higher opinion of the support system may lead to worse classification results.

5 Discussion

In this section, we discuss our results, answer our RQs from Section 1, and provide recommendations for the design of anti-phishing systems. In addition, we explain the limitations of our study.

5.1 Research Questions

5.1.1 RQ1: Effectiveness of Support Systems. We found no differences between the support systems and the control group in terms of the correct classification of all emails. When looking exclusively at the emails where the support system was displayed, only the warning banner showed an improvement compared to the control group. Similar to other studies [8, 22, 25, 42], we found that context plays a critical role for users in determining whether an email is legitimate or fraudulent. In our study, the email context and scenario had a greater impact on email classification than the actual support systems. When considering all emails in a logistic regression model (see Section 4.2.1), the contact list had a positive effect on the correct classification of emails. Regarding the calendar, the effect was minor and not significant, and the job offer document even had an adverse effect. Meanwhile, the support systems made no significant difference, except for the tooltip, which led to a slight decline. At the same time, we found that hovering over a link had a positive effect on the correct classification, as shown by the regression model. Meanwhile, URLs are often difficult for users to understand and interpret, making it unclear where they might lead. Other studies [5, 6, 35] address this issue and aim to support users with URL parsing. Torpedo [55] also includes special highlighting of the domain, which simultaneously increases readability to help users understand the URL.

At the same time, there might be other reasons why the support systems did not have much of an impact. We cover this in more detail in Section 5.3.

Furthermore, participants reported that they most frequently looked at relevant cues, namely the sender's email address and URLs, followed closely by the sender's name, which can be easily faked. In contrast, the sender's email address is better protected by technical features such as DKIM, SPF, and DMARC. The support systems played a secondary role here, as the participants mentioned them less frequently, and when they did, they ranked them lower in relevance. The warning banner was an exception and the third-most-often-mentioned item, and at the same time, it was also most often cited as conspicuous.

5.1.2 RQ2: Overcautious Behavior. Regarding RQ2, which examined whether support systems cause overcautious behaviour, we examined the number of legitimate emails our participants misclassified as phishing. The tooltip led to two legitimate emails being classified as phishing by more participants than in all other systems and the control group. However, we cannot generalize this, because the tooltip was only displayed for emails containing links, and our regression model shows that such emails were generally harder to classify. When comparing the classification of emails with false positives to the same emails in the control groups, we found no effects. Hence, our results remain inconclusive, and we cannot say that one particular support system caused overcautious behavior in our study.

5.1.3 RQ3: Influence of False Positives and False Negatives in Support System Classification. We found that false positives had no significant impact on participants' subsequent classifications. In the case of the support systems that we had integrated, false positives (label and tooltip), participants frequently criticized their reliability and accuracy compared to the other support systems. Participants also gave the label a lower rating regarding how much they trusted it and how much it affected their decisions, likely due to perceived unreliability.

A relatively small number of false positives have been identified, but these have led to the systems being criticized for their reliability. Thus, even a few false positives could lead to mistrust of support systems or alarm fatigue in the future, as support systems will never perform perfectly, and a perfect technical classification of phishing emails would make support systems obsolete. The fact that the false positives did not affect subsequent classifications could be explained by participants noticing the incorrect classifications and, consequently, becoming mistrustful of the support system.

Regarding false negatives, a difference was observed in the external marker after the false-negative email. This may be explained by the fact that participants already understood how the external marker worked after seeing it in a few emails. With other support systems, it is more difficult to understand their classification process and when they should occur in an email. In general, it is more difficult to detect false negatives than false positives, and it must be assumed that participants often did not notice the absence of a support system.

5.1.4 RQ4: Perception of Support Systems by Users. We can answer RQ4 by concluding that users appreciate the support systems and find them helpful, even though we found no impact on the correct classification of phishing emails. Users may perceive their effectiveness as greater than it actually is. Participants claimed that the warning banner, external marker, and tooltip influenced their decisions and that they trusted them. In contrast, the label received a neutral rating, suggesting a lower perceived reliability. Generally, the accuracy and reliability were criticized, particularly for the label, and to some extent for the tooltip. However, the tooltip was criticized more for providing an "unclear assessment", whereby the label for its high error rate. Although the participants uncritically followed the assessment of the support system a few times, more often they mentioned that support systems increased their awareness and encouraged them to check their emails more carefully. At the same time, participants expressed skepticism or uncertainty about the assessment of the support system, particularly regarding the label. On the other hand, we found limited evidence that participants felt protected by the support system or developed a false sense of security, as observed in other studies [16, 21, 61].

5.2 Recommendations for Anti-Phishing Interfaces

5.2.1 Utilizing Contextual Clues. We found that participants often paid attention to irrelevant phishing cues, such as the sender's name, and context had a stronger impact than the support systems themselves. It could be beneficial to use this fact and focus attention on contextual clues rather than relying solely on technical indicators. Especially in cases where the support system is "unsure" in its decision and provides an unclear assessment, it could instead encourage users to be aware of external, trustworthy information outside the digital environment. In organizations, collaborative tools that interconnect and share information are commonplace. Here, information from other applications could be integrated into the support systems, for example, that the sender is not yet in the contact list, or that an event is about to take place (or took place recently) according to the calendar.

5.2.2 Highlight Suspicious Elements. To draw users' attention to the relevant phishing cues, they should be emphasized without resorting to overly technical language. For example, discrepancies between the URL and the displayed text in the email could be pointed out. Unfortunately, this does not help when the link is hiding behind a button or other graphical element. Overall, as link hovering was found to be one of the most important factors in correct classification, systems that help users to interpret links seem to be a promising addition to anti-phishing systems.

Additionally, key terms indicating suspicious activity and requests to share sensitive data can be highlighted, such as requests for passwords, logins, and credentials.

5.2.3 Avoid unclear assessments. In our case, the tooltip often led participants to classify legitimate emails as phishing when its assessment was unclear. Moreover, systems with unclear assessments (tooltip and label) were perceived as unhelpful and untrustworthy more frequently than deterministic systems (external marker and warning banner), even though the latter exhibited false negatives. Support systems should make a clear statement; otherwise, they may confuse users, leading them to either become overly cautious or to rely on systems at all. Concrete indications of why the support system has made an assessment are more helpful than vague statements.

5.2.4 Balancing Conspicuousness and Over-Conspicuousness. Some support systems, especially the external marker and label, were often overlooked by participants. At the same time, we found that yellow is already very conspicuous in the warning banner. Red, for example, could signal excessive danger and lead users to become even more overcautious. Additionally, size and placement appear to be decisive factors in ensuring the support system is noticed, though we did not investigate this in depth. Support systems should be self-explanatory and comprehensible without instructions or additional guidance. However, within organizations, their existence and purpose should be clearly communicated to raise awareness. Otherwise, there is a risk that they will go unnoticed or be misinterpreted.

5.3 Limitations

We developed the tool as an online study system in which participants role-played. The tool restricted participants' interaction with the emails, as it only provided limited capabilities. At the same time, they were constrained to work within the limited, predetermined context of the role. Besides that, they were aware from the briefing at the beginning that the study was about phishing. Additionally, clickworkers may differ from the general population, particularly in terms of greater IT affinity. In addition, it is likely that there are differences in how they processed the emails in the scenario and in the strategies a real user would use to evaluate emails. Therefore, the responses may not be representative of less experienced users. Although the study included a substantial number of participants, some small effects may not have been detected because of the statistical power limitations.

We replicated the support systems as best we could, but had to make adjustments to integrate them into the tool. Regarding the tooltip, as mentioned in Section 3.2.4 before, we did not install the actual TORPEDO plugin; instead, we mimicked its functionality. That means that the actual plugin might perform differently. Regarding the label, we were unable to determine how the support system's classification mechanisms work and therefore had to display the false positives and false negatives according to the specified probabilistic procedure.

The ratio of emails in our scenario is unbalanced, with 3 phishing emails and 15 legitimate emails, which reduces the statistical power of our analyses. Especially in the case of phishing emails, a single incorrectly or correctly classified email can strongly influence the result. We tried to create a ratio of fraudulent to genuine emails as close to reality as possible, while avoiding overwhelming participants with too many emails, as attentiveness tends to decline over time. Yet the number of phishing emails remains high compared to the actual rate [27, 30]. At the same time, our small selection of phishing emails allows us to simulate only a few types and limits generalizability. Many participants correctly classified most emails, which may reflect ceiling effects. This may be partly because we included many legitimate emails that did not contain links or attachments and could therefore be readily ruled out as phishing.

6 Conclusion and Future Work

In this study, we took the first steps toward a comparative evaluation of various anti-phishing support systems using an experimental tool developed specifically for this purpose. To this end, 453 participants slipped into the role of an HR employee in our designed scenario and classified 18 emails as fraudulent or legitimate. Subsequently, they expressed their opinions on the support systems in a concluding survey. Our results show that there is no significant difference between the support systems and the control group. However, we found a slightly adverse effect of the tooltip support system compared to other support systems. It became apparent that the provided context had a greater impact than support systems. At the same time, participants appreciated the support systems and considered them helpful, even though we were unable to determine any actual effect. In future work, the support systems could be examined outside of an online survey in real environments and workplaces. In addition, the ratio of phishing emails to legitimate emails would be aligned with real-world conditions. At the same time, participants who are not clickworkers but, for example, ordinary employees, whose participation takes place in the context of their daily work routine rather than as a paid survey task, would demonstrate a different approach to email processing. This would also result in a more heterogeneous participant sample, with greater variation in prior knowledge and IT affinity. Most importantly, false positives and false negatives in classification should be included in the experimental setup, as we showed that they influence reliance on support systems, and their influence should be further investigated.

Acknowledgments

We thank Claudia Ceh and Sebastian Pahl for their support in the development of our interactive tool. The authors used Grammarly to revise text and to correct grammatical errors. This work was funded by the European Union under grant number ROF-SG20-3066-3-2-2.

References

[1] 2025. Identity-Centric Threats: The New Reality. https://esentire-dot-com-assets.s3.amazonaws.com/assets/resourcefiles/eSentire_Report_Identity-Centric-Threats.pdf

[2] Lawrence Abrams. 2025. *Fake "Security Alert" Issues on GitHub Use OAuth App to Hijack Accounts.* https://www.bleepingcomputer.com/news/security/fake-security-alert-issues-on-github-use-oauth-app-to-hijack-accounts/

[3] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in warningland: a large-scale field study of browser security warning effectiveness. In *USENIX Security Symposium.*

[4] Sara Albakry, Kami Vaniea, and Maria K. Wolters. 2020. What is this URL's Destination? Empirical Evaluation of Users' URL Reading. In *ACM CHI.* Honolulu HI USA. doi:10.1145/3313831.3376168

[5] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. 2021. I Don't Need an Expert! Making URL Phishing Features Human Comprehensible. In *ACM CHI.* doi:10.1145/3411764.3445574

[6] Kholoud Althobaiti, Kami Vaniea, and Serena Zheng. 2018. Faheem: Explaining URLs to people using a Slack bot. In *Symposium on Digital Behaviour Intervention for Cyber Security.* 1–8.

[7] Joseph Aneke, Carmelo Ardito, and Giuseppe Desolda. 2019. Designing an Intelligent User Interface for Preventing Phishing Attacks. In *IFIP Conference on Human-Computer Interaction.* doi:10.1007/978-3-030-46540-7_10

[8] Zinaida Benenson, Freya Gassmann, and Robert Landwirth. 2017. Unpacking Spear Phishing Susceptibility. In *FC.* Springer International Publishing, 610–627. doi:10.1007/978-3-319-70278-0_39

[9] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W. Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. 2013. Your attention please: designing security-decision UIs to make genuine risks harder to ignore. In *SOUPS.* doi:10.1145/2501604.2501610

[10] Robert L. Brennan and Dale J. Prediger. 1981. Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement* 41, 3 (1981), 687–699. doi:10.1177/001316448104100307

[11] Lina Brunken, Annalina Buckmann, Jonas Hielscher, and M. Angela Sasse. 2023. "To Do This Properly, You Need More Resources": The Hidden Costs of Introducing Simulated Phishing Campaigns. In *USENIX Security Symposium.*

[12] Gamze Canova, Melanie Volkamer, Clemens Bergmann, Roland Borza, Benjamin Reinheimer, Simon Stockhardt, and Ralf Tenberg. 2015. Learn To Spot Phishing URLs with the Android NoPhish App. In *IFIPAICT.* doi:10.1007/978-3-319-18500-2_8

[13] Deanna D Caputo, Shari Lawrence Pfleeger, Jesse D Freeman, and M Eric Johnson. 2013. Going spear phishing: Exploring embedded training and awareness. *IEEE Security & Privacy* 12, 1 (2013), 28–38.

[14] A. Carella, Murat Kotsoev, and T. Truta. 2017. Impact of Security Awareness Training on Phishing Click-Through Rates. In *IEEE International Conference on Big Data (Big Data).* IEEE, 4458–4466. doi:10.1109/BigData.2017.8258485

[15] Xiaowei Chen, Margault Sacré, Gabriele Lenzini, Samuel Greiff, Verena Distler, and Anastasia Sergeeva. 2024. The Effects of Group Discussion and Role-playing Training on Self-efficacy, Support-seeking, and Reporting Phishing Emails: Evidence from a Mixed-design Experiment. (Feb. 2024). doi:10.1145/3613904.3641943 arXiv:2402.11862 [cs].

[16] Dan Conway, Ronnie Taib, Mitch Harris, Kun Yu, Shlomo Berkovsky, and Fang Chen. 2017. A Qualitative Investigation of Bank Employee Experiences of Information Security and Phishing. In *SOUPS*. Santa Clara, CA. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/conway

[17] Ronald C. Dodge, Curtis Carver, and Aaron J. Ferguson. 2007. Phishing for user security awareness. *Computers & Security* 26, 1 (Feb. 2007), 73–80. doi:10.1016/j.cose.2006.10.009

[18] 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 *ACM CHI*. doi:10.1145/1357054.1357219

[19] William J. Gordon, Adam Wright, Ranjit Aiyagari, Leslie Corbo, Robert J. Glynn, Jigar Kadakia, Jack Kufahl, Christina Mazzone, James Noga, Mark Parkulo, Brad Sanford, Paul Scheib, and Adam B. Landman. 2019. Assessment of Employee Susceptibility to Phishing Attacks at US Health Care Institutions. *JAMA Network Open* 2, 3 (2019), e190393. doi:10.1001/jamanetworkopen.2019.0393

[20] Kristen Greene, Michelle Steves, and Mary Theofanos. 2018. No Phishing beyond This Point. *Computer* 51, 6 (June 2018). doi:10.1109/MC.2018.2701632

[21] Kristen Greene, Michelle Steves, Mary Theofanos, and Jennifer Kostick. 2018. User Context: An Explanatory Variable in Phishing Susceptibility. In *Workshop on Usable Security*. doi:10.14722/usec.2018.23016

[22] Frank L. Greitzer, Wanru Li, Kathryn B. Laskey, James Lee, and Justin Purl. 2021. Experimental Investigation of Technical and Human Factors Related to Phishing Susceptibility. *ACM Transactions on Social Computing* 4, 2 (June 2021), 1–48. doi:10.1145/3461672

[23] Matthew L. Hale, Rose F. Gamble, and Philip Gamble. 2015. CyberPhishing: A Game-Based Platform for Phishing Awareness Testing. In *IEEE Hawaii International Conference on System Sciences*. doi:10.1109/HICSS.2015.670

[24] Ayako A. Hasegawa, Naomi Yamashita, Mitsuaki Akiyama, and Tatsuya Mori. 2021. Why They Ignore English Emails: The Challenges of Non-Native Speakers in Identifying Phishing Emails. In *SOUPS*. https://www.usenix.org/conference/soups2021/presentation/hasegawa

[25] Doron Hillman, Yaniv Harel, and Eran Toch. 2023. Evaluating organizational phishing awareness training on an enterprise scale. *Computers & Security* 132 (Sept. 2023), 103364. doi:10.1016/j.cose.2023.103364

[26] Grant Ho, Ariana Mirian, Elisa Luo, Khang Tong, Euyhyun Lee, Lin Liu, Christopher A. Longhurst, Christian Dameff, Stefan Savage, and Geoffrey M. Voelker. 2025. Understanding the Efficacy of Phishing Training in Practice. In *2025 IEEE Symposium on Security and Privacy (SP)*. 37–54. doi:10.1109/SP61157.2025.00076

[27] Hornetsecurity. 2024. *Nearly Half a Billion Emails to Businesses Contain Malicious Content, Hornetsecurity Report Finds*. Hornetsecurity – Next-Gen Microsoft 365 Security. https://www.hornetsecurity.com/en/blog/cyber-security-report-2025-press-release/

[28] Hang Hu, Peng Peng, and Gang Wang. 2018. Towards Understanding the Adoption of Anti-Spoofing Protocols in Email Systems. In *IEEE SecDev*. doi:10.1109/SecDev.2018.00020

[29] K. Jansson and R. V. Solms. 2013. Phishing for Phishing Awareness. *Behaviour & Information Technology* 32, 6 (2013). doi:10.1080/0144929X.2011.632650

[30] Keepnet Labs. [n. d.]. *2025 Phishing Statistics: (Updated August 2025) - Keepnet*. Keepnet Labs Blog. https://keepnetlabs.com/blog/top-phishing-statistics-and-trends-you-must-know

[31] Kat Krol, Matthew Moroz, and M. Angela Sasse. 2012. Don't work. Can't work? Why it's time to rethink security warnings. In *International Conference on Risks and Security of Internet and Systems (CRiSIS)*. doi:10.1109/CRISIS.2012.6378951

[32] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Protecting people from phishing: the design and evaluation of an embedded training email system. In *ACM CHI*. doi:10.1145/1240624.1240760

[33] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason I. Hong. 2010. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology* 10 (2010), 1–31.

[34] Daniele Lain, Kari Kostiainen, and Srdjan Čapkun. 2022. Phishing in Organizations: Findings from a Large-Scale and Long-Term Study. doi:10.1109/SP46214.2022.9833766

[35] Daniele Lain, Yoshimichi Nakatsuka, Kari Kostiainen, Gene Tsudik, and Srdjan Capkun. 2025. URL inspection tasks: helping users detect phishing links in emails. In *USENIX Security Symposium*.

[36] Sourena Maroofi, Maciej Korczyński, Arnold Hölzel, and Andrzej Duda. 2021. Adoption of email anti-spoofing schemes: a large scale analysis. *IEEE Transactions on Network and Service Management* 18, 3 (2021), 3184–3196.

[37] Steven McElwee, George Murphy, and Paul Shelton. 2018. Influencing Outcomes and Behaviors in Simulated Phishing Exercises. In *IEEE SoutheastCon*. doi:10.1109/SECON.2018.8479109

[38] Justin Petelka, Yixin Zou, and Florian Schaub. 2019. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings. In *ACM CHI*. doi:10.1145/3290605.3300748

[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. In *SOUPS*. https://www.usenix.org/conference/soups2020/presentation/reinheimer

[40] Joshua Reynolds, Deepak Kumar, Zane Ma, Rohan Subramanian, Meishan Wu, Martin Shelton, Joshua Mason, Emily Stark, and Michael Bailey. 2020. Measuring Identity Confusion with Uniform Resource Locators. In *ACM CHI*. doi:10.1145/3313831.3376298

[41] Stefan A. Robila and James W. Ragucci. 2006. Don't Be a Phish: Steps in User Education. In *ACM SIGCSE*. doi:10.1145/1140124.1140187

[42] Katharina Schiller, Florian Adamsky, Christian Eichenmüller, Matthias Reimert, and Zinaida Benenson. 2024. Employees' Attitudes towards Phishing Simulations: "It's like when a child reaches onto the hot hob". In *ACM CCS*. doi:10.1145/3658644.3690212

[43] Martin Schrepp. 2023. User Experience Questionnaire Handbook (Version 11). https://www.ueq-online.org/Material/Handbook.pdf

[44] Martin Schrepp. 2018. *UEQ - User Experience Questionnaire*. https://www.ueq-online.org/

[45] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). 4, 6 (2017), 103. doi:10.9781/ijimai.2017.09.001

[46] Jasmin Schwab, Alexander Nußbaum, Anastasia Sergeeva, Florian Alt, and Verena Distler. 2024. What Makes Phishing Simulation Campaigns (Un)Acceptable? A Vignette Experiment on the Acceptance and Manipulation Intention Related to Phishing Simulation Campaigns. 4737715 (Feb. 2024). doi:10.2139/ssrn.4737715

[47] Steve Sheng, Bryant Magnien, P. Kumaraguru, A. Acquisti, L. Cranor, J. Hong, and Elizabeth Ferrall-Nunge. 2007. Anti-Phishing Phil: The Design and Evaluation of a Game that Teaches People not to Fall for Phish. In *SOUPS*. doi:10.1145/1280680.1280692

[48] Hossein Siadati, Sean Palka, Avi Siegel, and Damon McCoy. 2017. Measuring the Effectiveness of Embedded Phishing Exercises. In *USENIX Workshop on CSET*. USENIX Association. https://www.usenix.org/conference/cset17/workshop-program/presentation/siadatii

[49] Michelle Steves, Kristen Greene, and Mary Theofanos. 2020. Categorizing human phishing difficulty: a Phish Scale. *Journal of Cybersecurity* 6, 1 (09 2020). doi:10.1093/cybsec/tyaa009 tyaa009.

[50] Simon Stockhardt, Benjamin Reinheimer, M. Volkamer, P. Mayer, Alexandra Kunz, P. Rack, and D. Lehmann. 2016. Teaching Phishing-Security: Which Way is Best?. In *International Conference on ICT Systems Security and Privacy Protection (IFIP SEC 2016)*. doi:10.1007/978-3-319-33630-5_10

[51] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. 2009. Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In *USENIX Security Symposium*.

[52] Jan Tolsdorf, David Langer, and Luigi Lo Iacono. 2025. Phishing Susceptibility and the (In-)Effectiveness of Common Anti-Phishing Interventions in a Large University Hospital. In *ACM CCS*. 4334–4348. doi:10.1145/3719027.3765164

[53] Bill Toulas. 2022. *Hackers Now Use 'Sock Puppets' for More Realistic Phishing Attacks*. BleepingComputer. https://www.bleepingcomputer.com/news/security/hackers-now-use-sock-puppets-for-more-realistic-phishing-attacks/

[54] Kai Florian Tschakert and Sudsanguan Ngamsuriyaroj. 2019. Effectiveness of and user preferences for security awareness training methodologies. *Heliyon* 5, 6 (June 2019), e02010. doi:10.1016/j.heliyon.2019.e02010

[55] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. 2017. User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. *Computers & Security* 71 (2017), 100–113. doi:10.1016/j.cose.2017.02.004

[56] Melanie Volkamer, Martina Angela Sasse, and Franziska Boehm. 2020. Analysing Simulated Phishing Campaigns for Staff. 312–328. doi:10.1007/978-3-030-66504-3_19

[57] Rick Wash and Molly M. Cooper. 2018. Who Provides Phishing Training?: Facts, Stories, and People Like Me. In *ACM CHI*. doi:10.1145/3173574.3174066

[58] Rick Wash, Norbert Nthala, and Emilee Rader. 2021. Knowledge and Capabilities that Non-Expert Users Bring to Phishing Detection. In *SOUPS*. USENIX Association. https://www.usenix.org/conference/soups2021/presentation/wash

[59] Patrickson Weanquoi, J. Johnson, and Jinghua Zhang. 2018. Using a Game to Improve Phishing Awareness. *Journal of Cybersecurity Education, Research and Practice* 2018, 2 (2018). doi:10.62915/2472-2707.1040

[60] Zikai Alex Wen, Z. Lin, Rowena Chen, and E. Andersen. 2019. What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. In *ACM CHI*. doi:10.1145/3290605.3300338

[61] Emma J. Williams, Joanne Hinds, and Adam N. Joinson. 2018. Exploring susceptibility to phishing in the workplace. *International Journal of Human-Computer Studies* 120 (2018), 1–13. doi:10.1016/j.ijhcs.2018.06.004

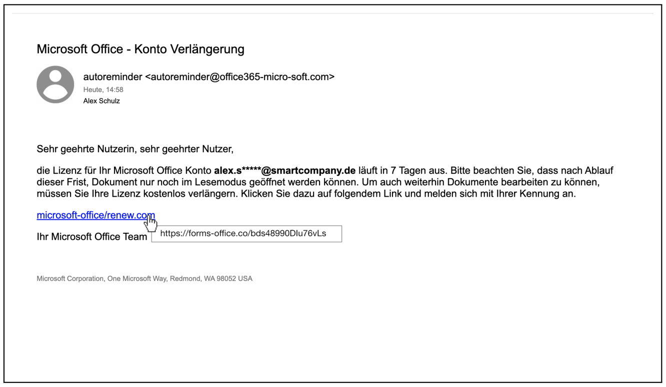
[62] William Yeoh, He Huang, Wang-Sheng Lee, Fadi Al Jafari, and Rachel Mansson. 2022. Simulated Phishing Attack and Embedded Training Campaign. *Journal of Computer Information Systems* 62, 4 (2022), 802–821. doi:10.1080/08874417.2021.1919941

[63] Sarah Y. Zheng and Ingolf Becker. 2023. Checking, nudging or scoring? Evaluating e-mail user security tools. In *SOUPS*. https://www.usenix.org/conference/soups2023/presentation/zheng

A Selected Emails

Below, the 3 selected phishing emails and the 15 selected legitimate emails are presented with screenshots and explanations. The screenshots are in German, but the explanations include English translations.

A.1 Phishing Emails



Microsoft Office - Konto Verlängerung

autoreminder <autoreminder@office365-micro-soft.com>
Heute, 14:58
Alex Schulz

Sehr geehrte Nutzerin, sehr geehrter Nutzer,

die Lizenz für Ihr Microsoft Office Konto **alex.s*****@smartcompany.de** läuft in 7 Tagen aus. Bitte beachten Sie, dass nach Ablauf dieser Frist, Dokument nur noch im Lesemodus geöffnet werden können. Um auch weiterhin Dokumente bearbeiten zu können, müssen Sie Ihre Lizenz kostenlos verlängern. Klicken Sie dazu auf folgendem Link und melden sich mit Ihrer Kennung an.

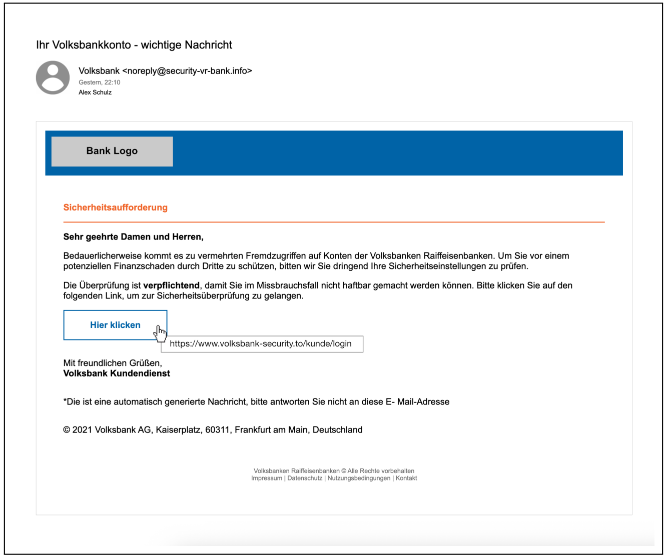
microsoft-office/renew.com

https://forms-office.co/bds48990Dlu76vLs

Ihr Microsoft Office Team

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052 USA

Figure 10: Phishing email "Microsoft": Subject: Microsoft Office - Account renewal. Body: Dear user, the license for your Microsoft Office account `alex.s*****@smartcompany.de` will expire in 7 days. Please note that after this period, documents can only be opened in read-only mode. To be able to continue editing documents, you must extend your license free of charge. To do this, click on the following link and log in with your ID. [[Link]] Your Microsoft Office Team



Ihr Volksbankkonto - wichtige Nachricht

Volksbank <noreply@security-vr-bank.info>
Gestern, 22:10
Alex Schulz

Sicherheitsaufforderung

Sehr geehrte Damen und Herren,

Bedauerlicherweise kommt es zu vermehrten Fremdzugriffen auf Konten der Volksbanken Raiffeisenbanken. Um Sie vor einem potenziellen Finanzschaden durch Dritte zu schützen, bitten wir Sie dringend Ihre Sicherheitseinstellungen zu prüfen.

Die Überprüfung ist **verpflichtend**, damit Sie im Missbrauchsfall nicht haftbar gemacht werden können. Bitte klicken Sie auf den folgenden Link, um zur Sicherheitsüberprüfung zu gelangen.

https://www.volksbank-security.to/kunde/login

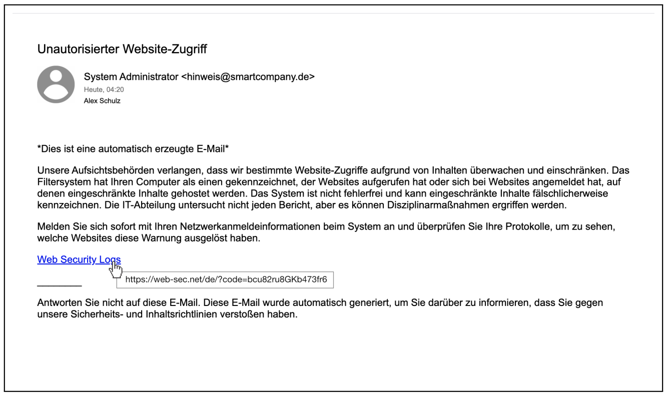
Mit freundlichen Grüßen,
Volksbank Kundendienst

*Die ist eine automatisch generierte Nachricht, bitte antworten Sie nicht an diese E- Mail-Adresse

© 2021 Volksbank AG, Kaiserplatz, 60311, Frankfurt am Main, Deutschland

Volksbanken Raiffeisenbanken © Alle Rechte vorbehalten
Impressum | Datenschutz | Nutzungsbedingungen | Kontakt

Figure 11: Phishing email "VR Bank": Subject: Your Volksbank Account – Important Message. Body: Security request - Dear Sir or Madam, unfortunately, there has been an increase in unauthorized access to the accounts of the Volksbanken Raiffeisenbanken. In order to protect you from potential financial losses caused by third parties, we urgently ask you to check your security settings. This check is mandatory so that you cannot be held liable in the event of misuse. Please click on the following link to access the security check. [[Link]] Yours sincerely, Volksbank Customer Service *This is an automatically generated message, please do not reply to this e-mail address.



Unautorisierter Website-Zugriff

System Administrator <hinweis@smartcompany.de>
Heute, 04:20
Alex Schulz

Dies ist eine automatisch erzeugte E-Mail

Unsere Aufsichtsbehörden verlangen, dass wir bestimmte Website-Zugriffe aufgrund von Inhalten überwachen und einschränken. Das Filtersystem hat Ihren Computer als einen gekennzeichnet, der Websites aufgerufen hat oder sich bei Websites angemeldet hat, auf denen eingeschränkte Inhalte gehostet werden. Das System ist nicht fehlerfrei und kann eingeschränkte Inhalte fälschlicherweise kennzeichnen. Die IT-Abteilung untersucht nicht jeden Bericht, aber es können Disziplinarmaßnahmen ergriffen werden.

Melden Sie sich sofort mit Ihren Netzwerkanmeldeinformationen beim System an und überprüfen Sie Ihre Protokolle, um zu sehen, welche Websites diese Warnung ausgelöst haben.

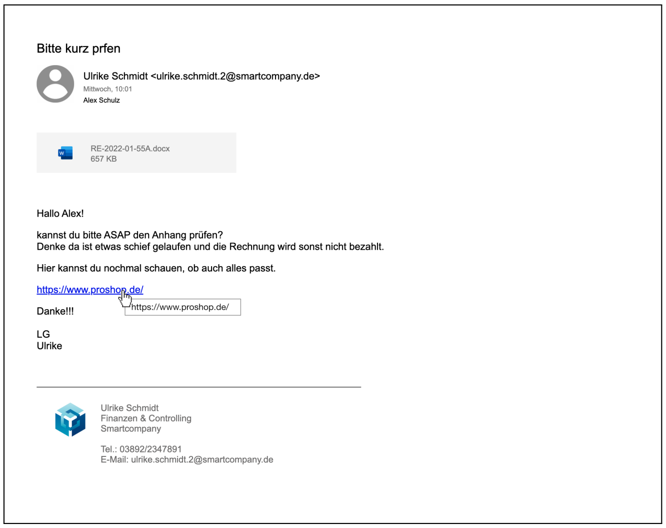
Web Security Logs

https://web-sec.net/de/?code=bcu82ru8GKb473fr6

Antworten Sie nicht auf diese E-Mail. Diese E-Mail wurde automatisch generiert, um Sie darüber zu informieren, dass Sie gegen unsere Sicherheits- und Inhaltsrichtlinien verstoßen haben.

Figure 12: Phishing email "Weblogs": Subject: Unauthorized Website Access. Body: *This is an automatically generated email* Our regulators require us to monitor and restrict certain website access based on content. The filtering system has flagged your computer as one that has accessed or logged into websites that host restricted content. The system is not error-free and may falsely flag restricted content. The IT department does not investigate every report, but disciplinary action may be taken. Log in to the system immediately with your network credentials and check your logs to see which websites triggered this warning. [[Link]] Do not respond to this email. This email has been automatically generated to inform you that you have violated our security and content policies.

A.2 Legitimate Emails



Bitte kurz prfen

Ulrike Schmidt <ulrike.schmidt.2@smartcompany.de>
Mittwoch, 10:01
Alex Schulz

RE-2022-01-55A.docx
657 KB

Hallo Alex!

kannst du bitte ASAP den Anhang prüfen?
Denke da ist etwas schief gelaufen und die Rechnung wird sonst nicht bezahlt.

Hier kannst du nochmal schauen, ob auch alles passt.

https://www.proshop.de/

https://www.proshop.de/

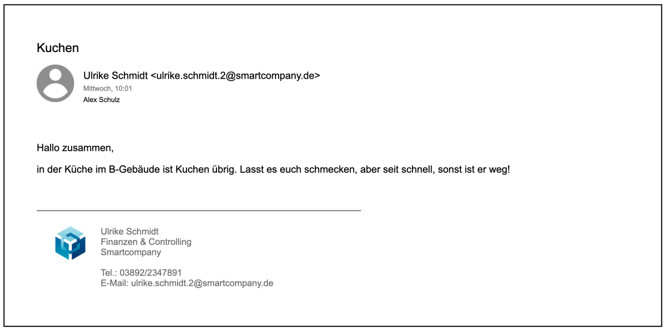
Danke!!!

LG
Ulrike

Ulrike Schmidt
Finanzen & Controlling
Smartcompany

Tel.: 03892/2347891
E-Mail: ulrike.schmidt.2@smartcompany.de

Figure 13: Legitimate email "Attachment": Subject: Please chck [sic] briefly. Body: Hi Alex! Can you please check the attachment ASAP? I think something has gone wrong and the invoice won't be paid otherwise. You can check here again to see if everything is OK. [[LINK]] Thanks! BR Ulrike



Kuchen

Ulrike Schmidt <ulrike.schmidt.2@smartcompany.de>
Mittwoch, 10:01
Alex Schulz

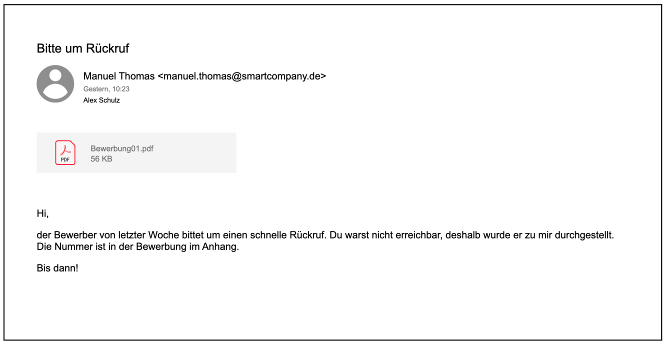
Hallo zusammen,

in der Küche im B-Gebäude ist Kuchen übrig. Lasst es euch schmecken, aber seit schnell, sonst ist er weg!

Ulrike Schmidt
Finanzen & Controlling
Smartcompany

Tel.: 03892/2347891
E-Mail: ulrike.schmidt.2@smartcompany.de

Figure 14: Legitimate email "Cake": Subject: Cake Body: Hello everyone, there is some cake left in the kitchen in Building B. Enjoy, but be quick, otherwise it will be gone!



Bitte um Rückruf

Manuel Thomas <manuel.thomas@smartcompany.de>
Gestern, 10:23
Alex Schulz

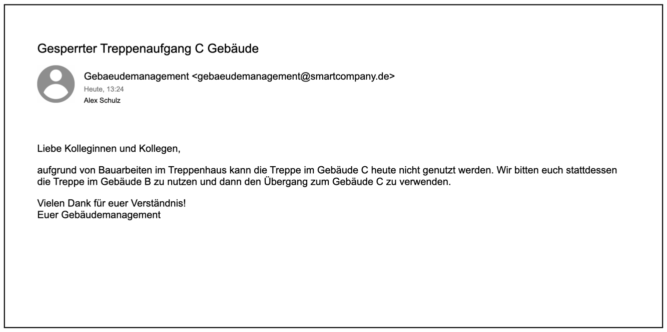
Bewerbung01.pdf
56 KB

Hi,

der Bewerber von letzter Woche bittet um einen schnelle Rückruf. Du warst nicht erreichbar, deshalb wurde er zu mir durchgestellt.
Die Nummer ist in der Bewerbung im Anhang.

Bis dann!

Figure 15: Legitimate email "Call back": Subject: Please call me back. Body: Hi, the applicant from last week is requesting a quick call back. You were unavailable, so the call was forwarded to me. The number is in the application attached. See you later!



Gesperrter Treppenaufgang C Gebäude

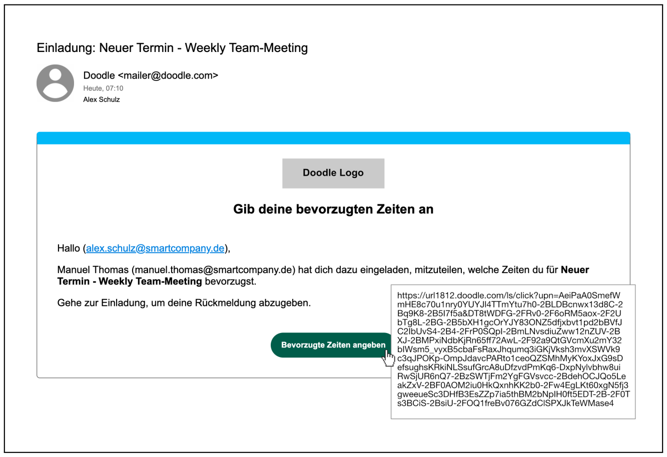
Gebaeudemanagement <gebaeudemanagement@smartcompany.de>
Heute, 13:24
Alex Schulz

Liebe Kolleginnen und Kollegen,

aufgrund von Bauarbeiten im Treppenhaus kann die Treppe im Gebäude C heute nicht genutzt werden. Wir bitten euch stattdessen die Treppe im Gebäude B zu nutzen und dann den Übergang zum Gebäude C zu verwenden.

Vielen Dank für euer Verständnis!
Euer Gebäudemanagement

Figure 16: Legitimate email "Construction": Subject: Closed stairway C building. Body: Dear colleagues, due to construction work in the stairwell, the stairs in Building C cannot be used today. We kindly ask you to use the stairs in Building B instead and then use the passageway to Building C. Thank you for your understanding! Your facility management team



Einladung: Neuer Termin - Weekly Team-Meeting

Doodle <mailer@doodle.com>
Heute, 07:10
Alex Schulz

Gib deine bevorzugten Zeiten an

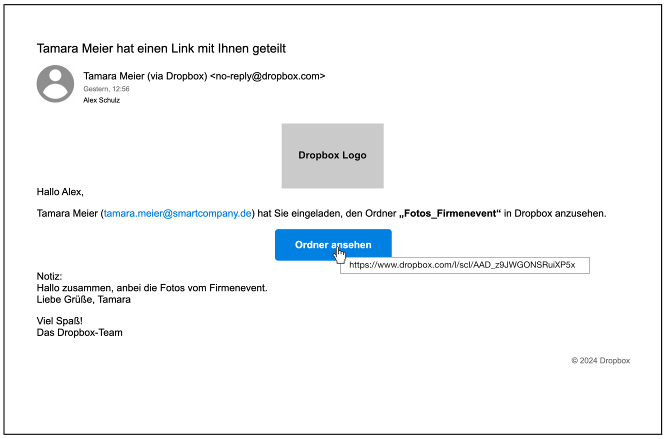
Hallo (alex.schulz@smartcompany.de),

Manuel Thomas (manuel.thomas@smartcompany.de) hat dich dazu eingeladen, mitzuteilen, welche Zeiten du für **Neuer Termin - Weekly Team-Meeting** bevorzugst.

Gehe zur Einladung, um deine Rückmeldung abzugeben.

https://url1812.doodle.com/ls/click?upn=AeiPaA0SmefWmHE8c70u1nry0YUYJl4TTmYtu7h0-2BLDBcnwx13d8C-2Bq9K8-2B5I7f5a&DT8tWDFG-2FRv0-2F6oRM5aox-2F2UbTg8L-2BG-2B5bXH1gcOrYJY83ONZ5dfjxbvt1pd2bBVfJC2IbUvS4-2B4-2FrP0SQpI-2BmLNvsdiuZww12nZUV-2BXJ-2BMPxiNdbKjRn65ff72AwL-2F92a9QtGVcmXu2mY32bIWsm5_vyxB5cbaFsRaxJhqumq3iGKjVksh3mvXSWVk9c3qJPOKp-OmpJdavcPARto1ceoQZSMhMyKYoxJxG9sDefsughsKRkiNLSsufGrcA8uDfzvdPmKq6-DxpNylvbhw8uiRwSjUR6nQ7-2BzSWTjFm2YgFGVsvcc-2BdehOCJQo5LeakZxV-2BF0AOM2iu0HkQxnhKK2b0-2Fw4EgLKt60xgN5fj3gweeueSc3DHfB3EsZZp7ia5thBM2bNpIH0ft5EDT-2B-2F0Ts3BCiS-2BsiU-2FOQ1freBv076GZdClSPXJkTeWMase4

Figure 17: Legitimate email "Doodle": Subject: Invitation: New Appointment - Weekly Team-Meeting. Body: Please indicate your preferred times. Hello (`alex.schulz@smartcompany.de`) Manuel Thomas (`manuel.thomas@smartcompany.de`) has invited you to share your preferred times for New appointment - Weekly team meeting. Go to the invitation to provide your feedback.



Tamara Meier hat einen Link mit Ihnen geteilt

Tamara Meier (via Dropbox) <no-reply@dropbox.com>
Gestern, 12:56
Alex Schulz

Hallo Alex,

Tamara Meier (tamara.meier@smartcompany.de) hat Sie eingeladen, den Ordner **„Fotos_Firmenevent“** in Dropbox anzusehen.

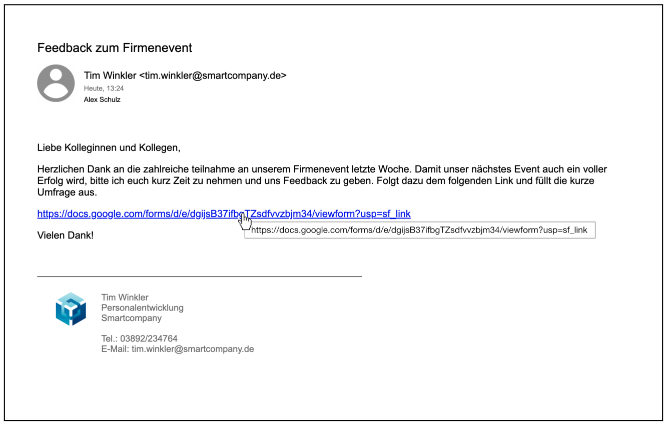
https://www.dropbox.com/l/scl/AAD_z9JWGONSRuiXP5x

Notiz:
Hallo zusammen, anbei die Fotos vom Firmenevent.
Liebe Grüße, Tamara

Viel Spaß!
Das Dropbox-Team

© 2024 Dropbox

Figure 18: Legitimate email "Dropbox": Subject: Tamara Meier shared a link with you. Body: Hello Alex, Tamara Meier has invited you to view the folder "Photos_CompanyEvent" in Dropbox. [[Button: View folder]]. Note: Hello everyone, please find attached the photos from the company event. Best regards, Tamara. Enjoy! The Dropbox Team.



Feedback zum Firmenevent

Tim Winkler <tim.winkler@smartcompany.de>
Heute, 13:24
Alex Schulz

Liebe Kolleginnen und Kollegen,

Herzlichen Dank an die zahlreiche teilnahme an unserem Firmenevent letzte Woche. Damit unser nächstes Event auch ein voller Erfolg wird, bitte ich euch kurz Zeit zu nehmen und uns Feedback zu geben. Folgt dazu dem folgenden Link und füllt die kurze Umfrage aus.

https://docs.google.com/forms/d/e/dgijsB37ifbgTZsdfvvzbjm34/viewform?usp=sf_link

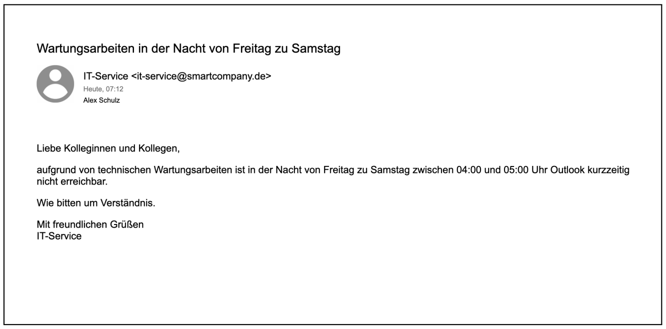
https://docs.google.com/forms/d/e/dgijsB37ifbgTZsdfvvzbjm34/viewform?usp=sf_link

Vielen Dank!

Tim Winkler
Personalentwicklung
Smartcompany

Tel.: 03892/234764
E-Mail: tim.winkler@smartcompany.de

Figure 19: Legitimate email "Feedback": Subject: Feedback on the Company Event. Body: Dear colleagues, Thank you very much for participating in our company event last week. To ensure that our next event is also a complete success, I would like to ask you to take a moment to provide us with some feedback. Please follow the link below and complete the short survey. [[Link]] Thanks!



Wartungsarbeiten in der Nacht von Freitag zu Samstag

IT-Service <it-service@smartcompany.de>
Heute, 07:12
Alex Schulz

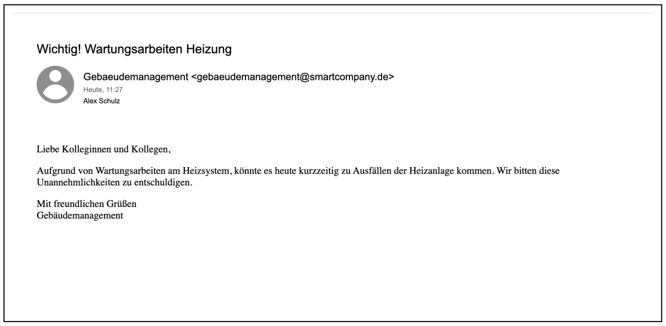
Liebe Kolleginnen und Kollegen,

aufgrund von technischen Wartungsarbeiten ist in der Nacht von Freitag zu Samstag zwischen 04:00 und 05:00 Uhr Outlook kurzzeitig nicht erreichbar.

Wie bitten um Verständnis.

Mit freundlichen Grüßen
IT-Service

Figure 20: Legitimate email "Service": Subject: Maintenance work during the night from Friday to Saturday. Body: Dear colleagues, due to technical maintenance work, Outlook will be temporarily unavailable during the night from Friday to Saturday between 04:00 and 05:00 a.m. We ask for your understanding. Yours sincerely, IT Service.



Wichtig! Wartungsarbeiten Heizung

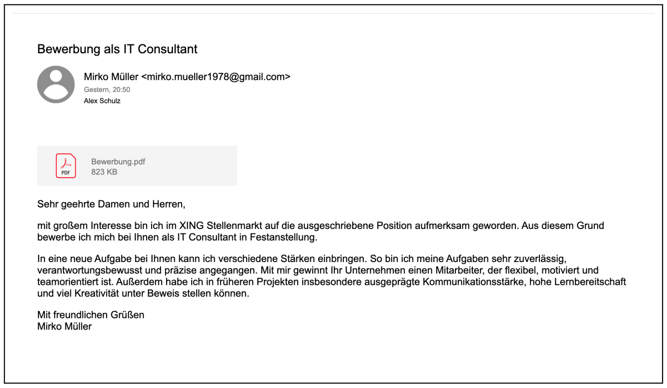
Gebaeudemanagement <gebaeudemanagement@smartcompany.de>
Heute, 11:27
Alex Schulz

Liebe Kolleginnen und Kollegen,

Aufgrund von Wartungsarbeiten am Heizsystem, könnte es heute kurzzeitig zu Ausfällen der Heizanlage kommen. Wir bitten diese Unannehmlichkeiten zu entschuldigen.

Mit freundlichen Grüßen
Gebäudemanagement

Figure 21: Legitimate email "Heating": Subject: Important! Heating Maintenance Work. Body: Dear colleagues, Due to maintenance work on the heating system, there may be brief interruptions to the heating today. We apologize for any inconvenience caused. Kind regards, Facility Management



Bewerbung als IT Consultant

Mirko Müller <mirko.mueller1978@gmail.com>
Gestern, 20:50
Alex Schulz

Bewerbung.pdf
823 KB

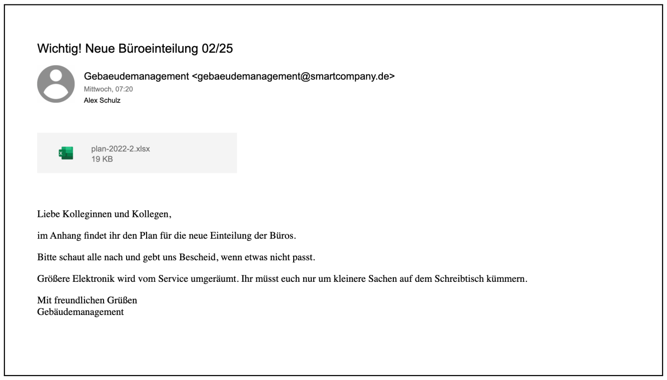
Sehr geehrte Damen und Herren,

mit großem Interesse bin ich im XING Stellenmarkt auf die ausgeschriebene Position aufmerksam geworden. Aus diesem Grund bewerbe ich mich bei Ihnen als IT Consultant in Festanstellung.

In eine neue Aufgabe bei Ihnen kann ich verschiedene Stärken einbringen. So bin ich meine Aufgaben sehr zuverlässig, verantwortungsbewusst und präzise angegangen. Mit mir gewinnt Ihr Unternehmen einen Mitarbeiter, der flexibel, motiviert und teamorientiert ist. Außerdem habe ich in früheren Projekten insbesondere ausgeprägte Kommunikationsstärke, hohe Lernbereitschaft und viel Kreativität unter Beweis stellen können.

Mit freundlichen Grüßen
Mirko Müller

Figure 22: Legitimate email "Job Application": Subject: Application as IT Consultant. Body. Dear Sir or Madam, I was very interested to see the advertised position on the XING job board. I am therefore applying for the permanent position of IT Consultant. I can contribute a variety of strengths to a new role with your company. I approach my tasks with a high degree of reliability, responsibility, and precision. With me, your company will gain an employee who is flexible, motivated, and team-oriented. In addition, I have demonstrated strong communication skills, a high willingness to learn, and a great deal of creativity in previous projects. Sincerely, Mirko Müller.



Wichtig! Neue Büroeinteilung 02/25

Gebaeudemanagement <gebaeudemanagement@smartcompany.de>
Mittwoch, 07:20
Alex Schulz

plan-2022-2.xlsx
19 KB

Liebe Kolleginnen und Kollegen,

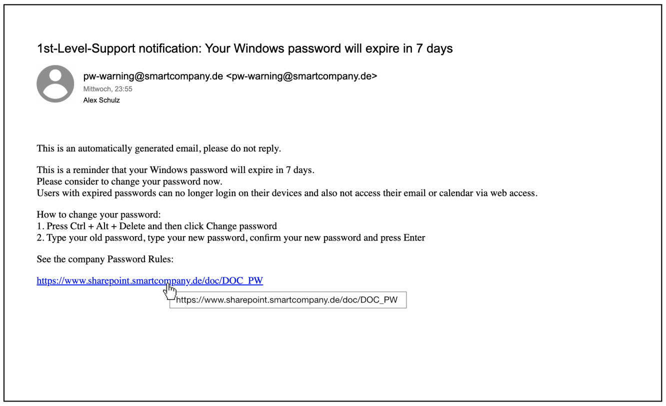
im Anhang findet ihr den Plan für die neue Einteilung der Büros.

Bitte schaut alle nach und gebt uns Bescheid, wenn etwas nicht passt.

Größere Elektronik wird vom Service umgeräumt. Ihr müsst euch nur um kleinere Sachen auf dem Schreibtisch kümmern.

Mit freundlichen Grüßen
Gebäudemanagement

Figure 23: Legitimate email "Office" Subject: Important! New office layout 02/25. Body: Dear colleagues, attached you will find the plan for the new office layout. Please take a look and let us know if anything is not right. Larger electronic equipment will be moved by the service department. You only need to take care of smaller items on your desk. Kind regards, Facility Management.



1st-Level-Support notification: Your Windows password will expire in 7 days

pw-warning@smartcompany.de <pw-warning@smartcompany.de>
Mittwoch, 23:55
Alex Schulz

This is an automatically generated email, please do not reply.

This is a reminder that your Windows password will expire in 7 days.
Please consider to change your password now.
Users with expired passwords can no longer login on their devices and also not access their email or calendar via web access.

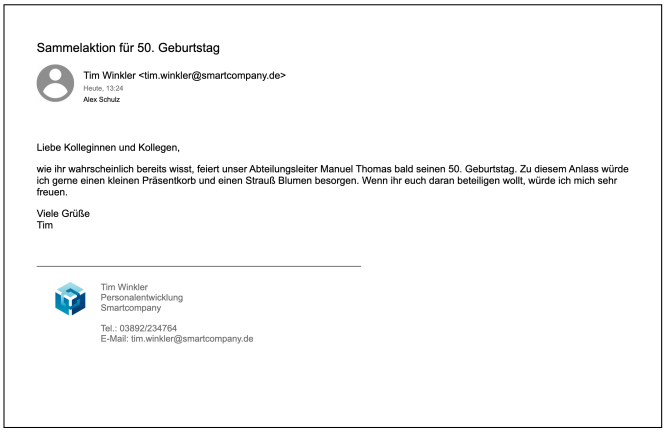
How to change your password:
1. Press Ctrl + Alt + Delete and then click Change password
2. Type your old password, type your new password, confirm your new password and press Enter

See the company Password Rules:

https://www.sharepoint.smartcompany.de/doc/DOC_PW

https://www.sharepoint.smartcompany.de/doc/DOC_PW

Figure 24: Legitimate email "Password" Subject: 1st-Level-Support notification: Your Windows password will expire in 7 days. Body: This is an automatically generated email, please do not reply. This is a reminder that your Windows password will expire in 7 days. Please consider to change your password now. Users with expired passwords can no longer login on their devices and also not access their email or calendar via web access. How to change your password: 1. Press Ctrl + Alt + Delete and then click Change password 2. Type your old password, type your new password, confirm your new password and press Enter See the company Password Rules: [[Link]].



Sammelaktion für 50. Geburtstag

Tim Winkler <tim.winkler@smartcompany.de>
Heute, 13:24
Alex Schulz

Liebe Kolleginnen und Kollegen,

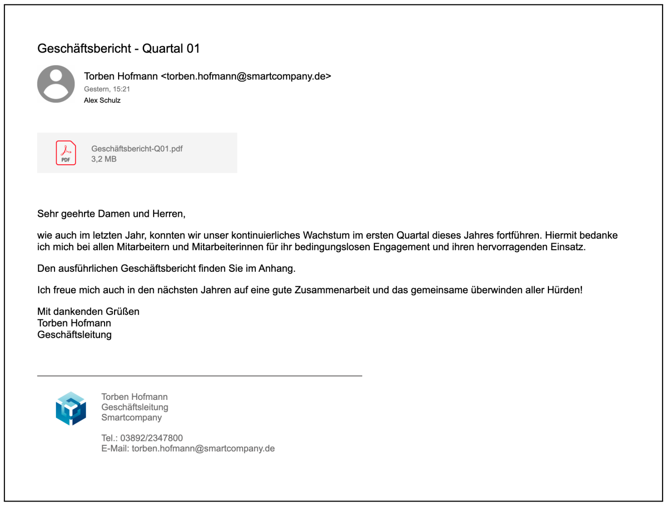
wie ihr wahrscheinlich bereits wisst, feiert unser Abteilungsleiter Manuel Thomas bald seinen 50. Geburtstag. Zu diesem Anlass würde ich gerne einen kleinen Präsentkorb und einen Strauß Blumen besorgen. Wenn ihr euch daran beteiligen wollt, würde ich mich sehr freuen.

Viele Grüße
Tim

Tim Winkler
Personalentwicklung
Smartcompany

Tel.: 03892/234764
E-Mail: tim.winkler@smartcompany.de

Figure 25: Legitimate email "Present". Subject: Fundraising campaign for 50th birthday. Body: Dear colleagues, as you probably already know, our head of department Manuel Thomas will soon be celebrating his 50th birthday. To mark the occasion, I would like to organize a small gift basket and a bouquet of flowers. If you would like to contribute, I would be very happy to do so. Best Regards Tim.



Geschäftsbericht - Quartal 01

Torben Hofmann <torben.hofmann@smartcompany.de>
Gestern, 15:21
Alex Schulz

Geschäftsbericht-Q01.pdf
3,2 MB

Sehr geehrte Damen und Herren,

wie auch im letzten Jahr, konnten wir unser kontinuierliches Wachstum im ersten Quartal dieses Jahres fortführen. Hiermit bedanke ich mich bei allen Mitarbeitern und Mitarbeiterinnen für ihr bedingungslosen Engagement und ihren hervorragenden Einsatz.

Den ausführlichen Geschäftsbericht finden Sie im Anhang.

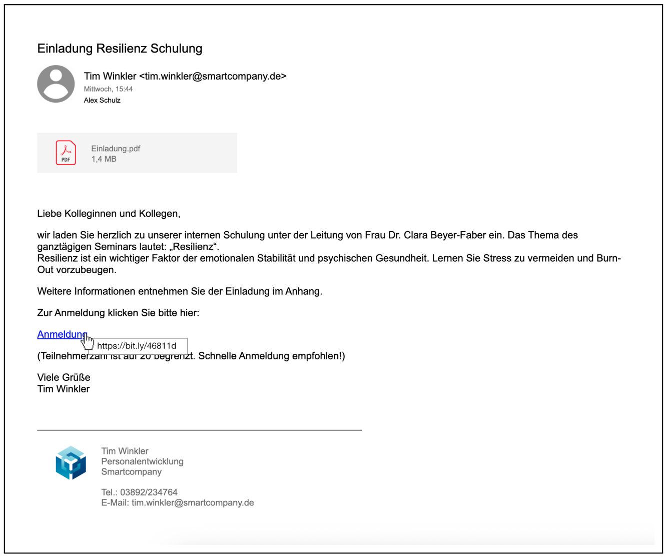
Ich freue mich auch in den nächsten Jahren auf eine gute Zusammenarbeit und das gemeinsame überwinden aller Hürden!

Mit dankenden Grüßen
Torben Hofmann
Geschäftsleitung

Torben Hofmann
Geschäftsleitung
Smartcompany

Tel.: 03892/2347800
E-Mail: torben.hofmann@smartcompany.de

Figure 26: Legitimate email "Report": Subject: Annual Report - Quarter 01. Body: Ladies and gentlemen, as in the previous year, we were able to continue our steady growth in the first quarter of this year. I would like to take this opportunity to thank all our employees for their unconditional commitment and outstanding efforts. You will find the detailed annual report attached. I look forward to continuing our good cooperation and overcoming all hurdles together in the coming years! Yours sincerely, Torben Hofmann, Managing Director.



Einladung Resilienz Schulung

Tim Winkler <tim.winkler@smartcompany.de>
Mittwoch, 15:44
Alex Schulz

Einladung.pdf
1,4 MB

Liebe Kolleginnen und Kollegen,

wir laden Sie herzlich zu unserer internen Schulung unter der Leitung von Frau Dr. Clara Beyer-Faber ein. Das Thema des ganztägigen Seminars lautet: „Resilienz".
Resilienz ist ein wichtiger Faktor der emotionalen Stabilität und psychischen Gesundheit. Lernen Sie Stress zu vermeiden und Burn-Out vorzubeugen.

Weitere Informationen entnehmen Sie der Einladung im Anhang.

Zur Anmeldung klicken Sie bitte hier:

Anmeldung

https://bit.ly/46811d

(Teilnehmerzahl ist auf 20 begrenzt. Schnelle Anmeldung empfohlen!)

Viele Grüße
Tim Winkler

Tim Winkler
Personalentwicklung
Smartcompany

Tel.: 03892/234764
E-Mail: tim.winkler@smartcompany.de

Figure 27: Legitimate email "Resilience": Subject: Invitation to Resilience Training. Body: Dear colleagues, we cordially invite you to our internal training course led by Dr. Clara Beyer-Faber. The topic of the full-day seminar is: "Resilience". Resilience is an important factor in emotional stability and mental health. Learn how to avoid stress and prevent burnout. Further information can be found in the attached invitation. Please click here to register: [[Link]] (The number of participants is limited to 20. Quick registration is recommended!) Best regards, Tim Winkler.

B Demographic Data

Table 5: Demographics

		Banner		External		Label		Tooltip		Control Group		Total	
		No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Gender	**female**	25	28.7	39	39.4	27	31.8	35	38.5	41	45.1	167	36.9
	male	61	70.1	59	59.6	58	68.2	56	61.5	49	53.9	283	62.5
	no answer	1	1.2	1	1.0	0	0.0	0	0.0	1	1.1	3	0.7
Age	**younger than 20 years**	3	3.5	1	1.0	2	2.4	0	0.0	3	3.3	9	2.0
	20-29 years	19	21.8	21	21.2	20	23.5	14	15.4	14	15.4	88	19.4
	30-39 years	25	28.7	36	36.4	18	21.2	27	29.7	27	29.7	133	29.4
	40-49 years	13	14.9	21	21.2	23	27.1	23	25.3	25	27.5	105	23.2
	50-59 years	18	20.7	9	9.1	15	17.7	11	12.1	9	9.9	62	13.7
	60-69 years	3	3.5	3	3.0	3	3.5	11	12.1	11	12.1	31	6.8
	70 years or older	0	0.0	1	1.0	0	0.0	1	1.1	0	0.0	2	0.4
	no answer	6	6.9	7	7.1	4	4.7	4	4.4	2	2.2	23	5.1
IT affine	**little affine**	6	6.9	7	7.1	4	4.7	2	2.2	11	12.1	30	6.6
	average affine	44	50.6	53	53.5	53	62.4	58	63.7	52	57.1	260	57.4
	strongly affine	36	41.4	38	38.4	27	31.8	31	34.1	27	29.7	159	35.1
	no answer	1	1.2	1	1.0	1	1.2	0	0.0	1	1.1	4	0.9
IT area	**no**	58	66.7	67	67.7	57	67.1	63	69.2	71	78.0	316	69.8
	yes	28	32.2	29	29.3	26	30.6	28	30.8	18	19.8	129	28.5
	no answer	1	1.2	3	3.0	2	2.4	0	0.0	2	2.2	8	1.8
School	**high school**	65	74.7	80	80.8	61	71.8	65	71.4	72	79.1	343	79.1
	middle school	17	19.5	18	18.2	19	22.4	23	25.3	18	19.8	95	19.8
	secondary school	4	4.6	1	1.0	4	4.7	3	3.3	1	1.1	13	1.1
	still in school	0	0.0	0	0.0	1	1.2	0	0.0	0	0.0	1	0.0
	other	1	1.2	0	0.0	0	0.0	0	0.0	0	0.0	1	0.0
Professional degree	**phd**	2	2.3	3	3.0	2	2.4	1	1.1	0	0.0	8	1.8
	master	20	23.0	32	32.3	19	22.4	23	25.3	29	31.9	123	27.2
	bachelor	14	16.1	17	17.2	16	18.8	15	16.5	13	14.3	75	16.6
	technical degree	4	4.6	7	7.1	6	7.1	7	7.7	4	4.4	28	6.2
	vocational	31	35.6	23	23.2	30	35.3	36	39.6	28	30.8	148	32.7
	no professional degree	3	3.5	1	1.0	1	1.2	1	1.1	1	1.1	7	1.6
	other	2	2.3	3	3.0	2	2.4	0	0.0	3	3.3	10	2.2
	no answer	11	12.6	13	13.1	9	10.6	8	8.8	13	14.3	54	11.9
Job	**employee**	53	60.9	52	52.5	41	48.2	53	58.2	55	60.4	254	56.1
	self employed	18	20.7	22	22.2	19	22.4	19	20.9	19	20.9	97	21.4
	university student	5	5.8	15	15.2	12	14.1	5	5.5	5	5.5	42	9.3
	trainee	4	4.6	0	0.0	1	1.2	1	1.1	3	3.3	9	2.0
	student	0	0.0	2	2.0	2	2.4	0	0.0	0	0.0	4	0.9
	housewife	2	2.3	0	0.0	2	2.4	2	2.2	2	2.2	8	1.8
	retired	4	4.6	2	2.0	3	3.5	4	4.4	3	3.3	16	3.5
	unemployed	1	1.2	1	1.0	3	3.5	6	6.6	2	2.2	13	2.9
	other	0	0.0	3	3.0	1	1.2	0	0.0	0	0.0	4	0.9
	no answer	0	0.0	2	2.0	1	1.2	1	1.1	2	2.2	6	1.3

C Additional Results

C.1 Additional Experimental Results

Table 6: The table shows the results of the pairwise post hoc MWU tests, exclusively for medium and difficult emails, according to the classification from the preliminary study. p is Bonferroni-Holm corrected. Mean and median refer to the correctly classified emails.

Group 1	Group 2	n_1	n_2	Mean_1	Mean_2	Median_1	Median_2	U	p	significant	$\|r\|$	effect
External	Label	99	85	77.5 %	79.5 %	71.4 %	85.7 %	3804.0	0.985	-	0.09	negligible
External	Banner	99	87	77.5 %	77.3 %	71.4 %	85.7 %	4149.0	1.000	-	0.03	negligible
External	Tooltip	99	91	77.5 %	70.8 %	71.4 %	71.4 %	5338.5	0.188	-	0.16	small
External	Control Group	99	91	77.5 %	75.4 %	71.4 %	71.4 %	4812.0	1.000	-	0.06	negligible
Label	Banner	85	87	79.5 %	77.3 %	85.7 %	85.7 %	3939.0	1.000	-	0.06	negligible
Label	Tooltip	85	91	79.5 %	70.8 %	85.7 %	71.4 %	4877.0	0.021	⋆	0.23	small
Label	Control Group	85	91	79.5 %	75.4 %	85.7 %	71.4 %	4531.5	0.290	-	0.15	small
Banner	Tooltip	87	91	77.3 %	70.8 %	85.7 %	71.4 %	4779.0	0.125	-	0.18	small
Banner	Control Group	87	91	77.3 %	75.4 %	85.7 %	71.4 %	4416.0	0.828	-	0.10	small
Tooltip	Control Group	91	91	70.8 %	75.4 %	71.4 %	71.4 %	3622.5	0.801	-	0.11	small

Table 7: The table shows the results of pairwise post hoc MWU tests for legitimate emails as part of robustness checks. Only participants from the label group with exactly two false positives were considered, while all participants from the other groups were included. p-values are Bonferroni-Holm corrected. Mean refers to the correctly classified emails.

Group 1	Group 2	n_1	n_2	Mean_1	Mean_2	U	p	significant	$\|r\|$	effect
Banner	External	87	99	88.8	89.2	4284.5	1.000	-	0.00	negligible
Banner	Control Group	87	91	88.8	86.9	4389.5	1.000	-	0.10	negligible
Banner	Tooltip	87	91	88.8	83.4	4793	0.120	-	0.19	small
Banner	Label	87	19	88.8	88.4	820	1.000	-	0.01	negligible
External	Control Group	99	91	89.2	86.9	5029	1.000	-	0.10	small
External	Tooltip	99	91	89.2	83.4	5508.5	0.069	-	0.20	small
External	Label	99	19	89.2	88.4	941.5	1.000	-	0.00	negligible
Control Group	Tooltip	91	91	86.9	83.4	4630	1.000	-	0.10	small
Control Group	Label	91	19	86.9	88.4	773	1.000	-	0.07	negligible
Tooltip	Label	91	19	83.4	88.4	681	1.000	-	0.14	small

Table 8: The table shows the results of pairwise post hoc MWU tests for legitimate emails as part of robustness checks. Only participants from the label group with exactly one false positive were considered, while all participants from the other groups were included. p-values are Bonferroni-Holm corrected. Mean refers to the correctly classified emails.

Group 1	Group 2	n_1	n_2	Mean_1	Mean_2	U	p	significant	$\|r\|$	effect
Banner	External	87	99	88.8	89.2	4284.5	1.000	-	0.00	negligible
Banner	Control Group	87	91	88.8	86.9	4389.5	1.000	-	0.10	negligible
Banner	Tooltip	87	91	88.8	83.4	4793	0.120	-	0.19	small
Banner	Label	87	25	88.8	85.1	1256	1.000	-	0.11	small
External	Control Group	99	91	89.2	86.9	5029	1.000	-	0.10	small
External	Tooltip	99	91	89.2	83.4	5508.5	0.069	-	0.20	small
External	Label	99	25	89.2	85.1	1436	1.000	-	0.11	small
Control Group	Tooltip	91	91	86.9	83.4	4630	1.000	-	0.10	small
Control Group	Label	91	25	86.9	85.1	1198.5	1.000	-	0.04	negligible
Tooltip	Label	91	25	83.4	85.1	1074	1.000	-	0.04	negligible

Table 9: The table shows the results of pairwise post hoc MWU tests for phishing emails as part of robustness checks. Only participants from the label group with no false negatives were considered, while all participants from the other groups were included. p-values are Bonferroni-Holm corrected. Mean refers to the correctly classified emails.

Group 1	Group 2	n_1	n_2	Mean_1	Mean_2	U	p	significant	$\|r\|$	effect
Banner	External	87	99	85.8	82.8	4591.5	1.000	-	0.07	negligible
Banner	Control Group	87	91	85.8	83.2	4157	1.000	-	0.05	negligible
Banner	Tooltip	87	91	85.8	86.4	3771.5	1.000	-	0.05	negligible
Banner	Label	87	57	85.8	87.7	2331	1.000	-	0.06	negligible
External	Control Group	99	91	82.8	83.2	4438	1.000	-	0.01	negligible
External	Tooltip	99	91	82.8	86.4	4010.5	1.000	-	0.11	small
External	Label	99	57	82.8	87.7	2475	1.000	-	0.12	small
Control Group	Tooltip	91	91	83.2	86.4	3754	1.000	-	0.10	negligible
Control Group	Label	91	57	83.2	87.7	2317	1.000	-	0.11	small
Tooltip	Label	91	57	86.4	87.7	2560	1.000	-	0.01	negligible

Table 10: The table shows the results of pairwise post hoc MWU tests for phishing emails as part of robustness checks. Only participants from the label group with exactly one false negative were considered, while all participants from the other groups were included. p-values are Bonferroni-Holm corrected. Mean refers to the correctly classified emails.

Group 1	Group 2	n_1	n_2	Mean_1	Mean_2	U	p	significant	$\|r\|$	effect
Banner	External	87	99	85.8	82.8	4591.5	1.000	-	0.07	negligible
Banner	Control Group	87	91	85.8	83.2	4157	1.000	-	0.05	negligible
Banner	Tooltip	87	91	85.8	86.4	3771.5	1.000	-	0.05	negligible
Banner	Label	87	28	85.8	96.4	933.5	0.144	-	0.22	small
External	Control Group	99	91	82.8	83.2	4438	1.000	-	0.01	negligible
External	Tooltip	99	91	82.8	86.4	4010.5	0.706	-	0.11	small
External	Label	99	28	82.8	96.4	973.5	0.038	*	0.26	small
Control Group	Tooltip	91	91	83.2	86.4	3754	0.938	-	0.10	negligible
Control Group	Label	91	28	83.2	96.4	919.5	0.057	-	0.25	small
Tooltip	Label	91	28	86.4	96.4	1048.5	0.398	-	0.17	small

C.2 Additional Survey Results

C.2.1 Ranking of Phishing Cues. Figure 28 shows the ranking of the different phishing cues regarding focus in general (left) and in our study (right).

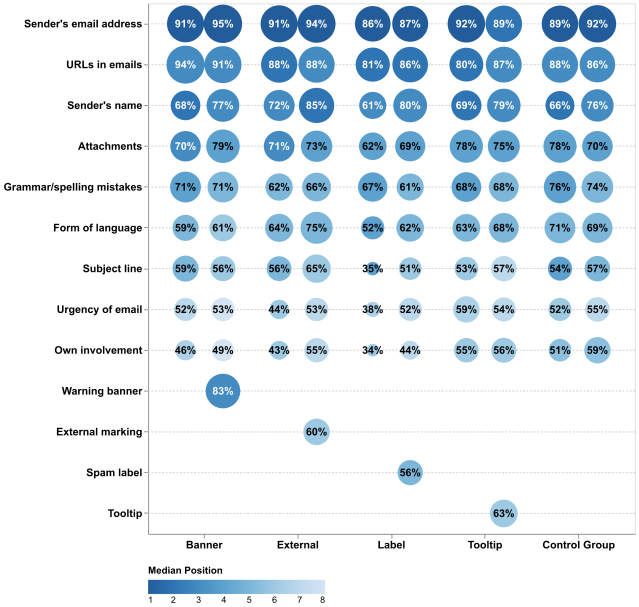


Figure 28: Ranking of the different phishing cues regarding focus in general (left) and in our study (right). The size of the bubbles and labels indicate how often participants mentioned an element within each support system group. The color represents the median position in which the element was ranked.

D Survey

Thank you very much! We have a few more questions for you. Please take the time to answer them.

(1) **Which parts of the emails did you pay the most attention to IN THE SURVEY?**
Arrange the answers in the list on the right that apply to you. Arrange the answers according to their relevance (highest relevance at the top). Use your mouse to move the answers. You do not have to select all answers. *Answer options in random order for each participant*
- Attachments
- Form of language
- Grammar/spelling mistakes
- Own involvement
- Sender's email address
- Sender's name
- Subject line
- Urgency of email
- URLs (Links, Web addresses) in the emails
- *Add own answer*

(2) **Which parts do you GENERALLY pay the most attention to when you receive a suspicious email?**
Arrange the answers in the list on the right that apply to you. Arrange the answers according to their relevance (highest relevance at the top). Use your mouse to move the answers. You do not have to select all answers. *Answer options in random order for each participant*
- Attachments
- Form of language
- Grammar/spelling mistakes
- Own involvement
- Sender's email address
- Sender's name
- Subject line
- Urgency of email
- URLs (Links, Web addresses) in the emails
- *Support system* (according to the test group: warning banner, external marker, spam label or tooltip (hint displayed for links))
- *Add own answer*

The following part was only shown to participants in the support system groups.

(3) **Have you noticed *the support system*? (Image)**
- Yes
- partially
- No

(4) **When you saw *the support system*, what was your first reaction?**
Only displayed to participants who have previously indicated yes or partially.

(5) **Please state to what extent you agree/disagree with the following statements.**
5-point Likert scale from strongly disagree to strongly agree
- I found *the support system* helpful.
- *The support system* impacted my decisions.
- I trusted *the support system.*
- I would also use *the support system* in my work life.
- I would also use *the support system* in my personal life.

(6) **Please give your assessment of *the support system*. The questionnaire consists of pairs of contrasting attributes that may apply to *the support system*.** Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression. Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely to *the support system.* Nevertheless, please tick a circle in every line.

obstructive	o	o	o	o	o	o	o	supportive
complicated	o	o	o	o	o	o	o	easy
inefficient	o	o	o	o	o	o	o	efficient
confusing	o	o	o	o	o	o	o	clear
boring	o	o	o	o	o	o	o	exciting
not interesting	o	o	o	o	o	o	o	interesting
conventional	o	o	o	o	o	o	o	inventive
usual	o	o	o	o	o	o	o	leading edge

(7) **Please rate your overall impression on a scale of 1–10:**
(1 = I don't like it at all, 10 = I like it very much)

(8) **Please justify your answer.**

The following part was displayed to all participants again.

(9) **Which gender do you feel you belong to?**
- Female
- Male
- Diverse
- No comment

(10) **What is your year of birth? Please select a year:** *selection list*

(11) **What is your current occupation?**
- Student
- Trainee
- University student
- Employee
- Self employed
- Unemployed
- Housewife/parental leave
- Retired
- Other
- No comment

(12) **What is your level of education?** (Please select the highest level of education you have achieved to date.)
- No general school leaving certificate
- Still in school
- Secondary school
- Middle school
- High school
- Other
- No comment

(13) **What is your professional degree?** (Please select the highest professional degree you have achieved to date.)
- No professional degree

- Vocational degree
- Technical degree
- Bachelor
- Master
- PhD
- Other
- No comment

(14) **Are you/were you working in the IT sector, or do you have prior knowledge (e.g., through a degree in computer science or similar) in this field?**

- Yes
- No
- No comment

(15) **How IT affine would you describe yourself?**

- Not very IT affine
- Moderately IT affine
- Very IT affine
- No comment

(16) **Which email client or web application do you mainly use PRIVATELY to read/reply to your emails?**

- Gmail
- Outlook Online (OWA)
- Outlook 365
- Outlook 2013
- Outlook 2010
- Apple Mail
- Mozilla Thunderbird
- I don't know
- Other
- No comment

(17) **Which email client or web application do you mainly use for WORK to read/reply to your emails?**

- Gmail
- Outlook Online (OWA)
- Outlook 365
- Outlook 2013
- Outlook 2010
- Apple Mail
- Mozilla Thunderbird
- I don't know
- Other
- No comment

(18) **Do you have any further comments? (Optional)**