

An investigation of phishing awareness and education over time: When and how to best remind users

Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, and Reyhan Duezguen, *SECUSO - Security, Usability, Society, Karlsruhe Institute of Technology*; Bettina Lofthouse, *Landesamt für Geoinformation und Landesvermessung Niedersachsen*; Tatiana von Landesberger, *Interactive Graphics Systems Group, Technische Universität Darmstadt*; Melanie Volkamer, *SECUSO - Security, Usability, Society, Karlsruhe Institute of Technology*

<https://www.usenix.org/conference/soups2020/presentation/reinheimer>

This paper is included in the Proceedings of the Sixteenth Symposium on Usable Privacy and Security.

August 10–11, 2020

978-1-939133-16-8

Open access to the Proceedings of the Sixteenth Symposium on Usable Privacy and Security is sponsored by USENIX.

An investigation of phishing awareness and education over time: When and how to best remind users

Benjamin Reinheimer*, Lukas Aldag*, Peter Mayer*, Mattia Mossano*, Reyhan Duezguen*,

Bettina Lofthouse†, Tatiana von Landesberger°, Melanie Volkamer*

* SECUSO - Security, Usability, Society, Karlsruhe Institute of Technology

† Landesamt für Geoinformation und Landesvermessung Niedersachsen

° Interactive Graphics Systems Group, Technische Universität Darmstadt

* *firstname.lastname@kit.edu*, ° *tatiana.von.landesberger@gris.tu-darmstadt.de*

† *Bettina.Lofthouse@lgln.niedersachsen.de*

Abstract

Security awareness and education programmes are rolled out in more and more organisations. However, their effectiveness over time and, correspondingly, appropriate intervals to remind users' awareness and knowledge are an open question. In an attempt to address this open question, we present a field investigation in a German organisation from the public administration sector. With overall 409 employees, we evaluated (a) the effectiveness of their newly deployed security awareness and education programme in the phishing context over time and (b) the effectiveness of four different reminder measures – administered after the initial effect had worn off to a degree that no significant improvement to before its deployment was detected anymore. We find a significantly improved performance of correctly identifying phishing and legitimate emails directly after and four months after the programme's deployment. This was not the case anymore after six months, indicating that reminding users after half a year is recommended. The investigation of the reminder measures indicates that measures based on videos and interactive examples perform best, lasting for at least another six months.

1 Introduction

Maintaining information security is an important challenge for organisations, and also for governmental and public administration sectors. The so-called German national IT Planning Council [71] requires German organisations in the public administration sector to implement information security management systems (ISMS). One of the goals of such ISMSs

is to enhance employees' information security awareness and knowledge. A common approach to satisfying this requirement is to roll out *security awareness and education programmes*. They typically raise general security awareness (e.g., everyone can potentially become a victim, the technological protection mechanisms need users' support, potential consequences of successful attacks) and convey knowledge about information security (including how to identify various attacks, how to reduce one's risks of becoming a victim of cyber attackers, and who shall be contacted in case of questions and incidents). These programmes may include *security awareness and education measures* that cover different aspects and/or topics using different media types, such as self-learning measures [63], e-learning platforms [9], on-site tutorials [102], or games [94]. Although such measures are widely deployed, an evaluation of their effectiveness related to their ability to enhance employees' information security skills over an extended time period is often missing. This, however, is of the essence: if employees are never, or only rarely, confronted with attacks that are included in a security awareness and education programme, the acquired awareness and knowledge might dissipate over time, as is the case with any other awareness and knowledge programmes. While waning of awareness and dwindling of knowledge is to be expected, it poses a problem to the maintenance of organisational information security. Therefore, it is crucial to know: (a) *when* awareness and knowledge levels should be renewed, i.e., how long the effect of a security awareness and education programme can be expected to last, and (b) *which* type of measures are best suited to restore users' awareness and knowledge.

Researchers could show the effectiveness of security awareness and education measures directly after roll-out [4, 64, 94, 98, 102, 106, 111, 119] and that the significant improvements endured over different spans of time. [22, 60, 116]. However, what is missing, is an insight into how long the impact of a security awareness and education measure lasts and how awareness and knowledge should be renewed. In order to gain these insights, we adopted and evaluated the phishing aware-

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Virtual Conference.

ness and education measure from [84]. We customised it for a German organisation from the public administration sector: a German State Office for Geoinformation and State Survey (SOGSS). We replaced example messages with those more suited to SOGSS and by removing irrelevant content. The content was presented face-to-face in on-site tutorials. A ‘*train the instructor*’ approach was used, which involved eleven instructors being trained by the Chief Information Security Manager. The participation in the tutorial was mandatory for all employees.

We evaluated employees’ skills in distinguishing phishing emails from legitimate emails at several points in time. First, data was collected just before and directly after the on-site tutorials. To study how long the effect lasted, we collected retention data four months after employees had participated in the tutorial. We were prepared to continue doing so every second month as long as we continued to see a significant enhancement of participants’ skills.

Our first contribution: we systematically measured the retention for the tutorial. Compared to previous studies we measured until the significant improvement wore off plus another measure after this point in time. The impact wore off after six months.

Our second contribution: we developed suitable reminder measures to replenish the employees’ phishing awareness and knowledge after receiving not significant results for the first time after the tutorial. We developed four different ones (three presenting the content using text, one using video and one using interactive email examples). The success of the reminder measures was evaluated right after their deployment, and again after six months.

Our third contribution: we accompanied an organization for a total of twelve months to both check for the effectiveness of the tutorial and the reminder measures. The awareness and knowledge levels of participants having either seen the video measure or the interactive examples after six months were still significantly higher twelve months after the initial tutorial.

As a consequence, SOGSSs decided to use the video and the interactive examples measures and to distribute these to all employees on a regular basis at six month intervals.

2 Related Work

This section commences by providing phishing definitions from the research literature. Related work is discussed next with regard to different types of security interventions, different study designs used to evaluate these interventions, and different types of tested users groups. Finally research into the impact of phishing security awareness and education measures over time are discussed.

Phishing Definitions: there are many different definitions of phishing in the literature. Correspondingly, researchers’ focus is different: (1) those who focus on phishers who want their victims to provide sensitive information (e.g. passwords,

bank details) using an authentic-looking phishing web page [1–4, 7, 16–19, 21, 22, 30, 33, 35, 38, 51, 54, 57, 58, 64, 68–70, 75–77, 80, 82, 88, 91, 93, 94, 96, 98, 104, 105, 108, 113–115, 117], or (2) those who focus on phishers who distribute malware when recipients click on links in messages or open attached files [1, 6, 8, 9, 12–15, 20, 23, 24, 31, 32, 34, 41, 43, 45, 48–50, 55, 56, 59–62, 74, 79, 81, 85–87, 90, 92, 99, 102, 103, 106, 107, 109–112, 116, 119]. Because it is safer to check the URL before clicking on it, instead of only checking the URL after opening the web page people are redirected to, *we focus on the second one*.

*Types of Interventions*¹: various studies evaluating different types of interventions to test their effectiveness exist. Researchers evaluated a range of tools that are supposed to provide further support (e.g. additional security indicators or displaying existing security indicators in different ways) [2, 5, 18, 29, 36, 40, 66, 68, 69, 72, 88, 96, 114, 115, 117, 118]. Different evaluated security awareness and education measures are a range of videos [46, 107], games [10, 11, 21, 22, 64, 94], various on-site instructor based tutorials [98, 102, 116] – as studied in the research presented in this paper – and a multitude of text-based measures [4, 47, 60, 65, 84, 92, 98, 102, 104, 110, 116, 119]. Additionally, there is research evaluating users’ skills in phishing detection without any interventions [6, 7, 12, 14, 15, 26, 30, 31, 34, 37, 39, 45, 48–51, 53, 54, 59, 73, 74, 80, 81, 85–87, 89, 90, 95, 100, 103, 109] (e.g. to understand decision making, to identify a baseline, or to motivate further research).

Study Designs: various types of lab studies have been employed, some with a cover story [4, 14, 15, 35, 36, 61, 69, 86, 91] and others without one [7, 11, 37, 40, 98, 103, 116], i.e., having security as participants’ primary goal by telling participants the goal of the user study. A number of remote studies have been carried out, including various types of online surveys, with phishing messages sent to the study participants own email accounts (not study specific) [30, 37, 40, 59, 81, 87, 102, 104], as well as to remotely accessible study-specific accounts [88, 90, 109, 110, 119]. Surveys include those that (1) show screenshots to be judged either as phishes or legitimate [54, 70, 100] as we did in our study. In some cases real phishes were used; others used examples created by the researchers; and (2) online surveys asking general questions such as the definition of phishing and the existing attack types [20, 51, 52, 79].

Types of User Groups: studies have targeted different user groups, i.e. mixed groups on a variety of panels without deliberately isolating specific kinds of participants [16, 35, 40, 54, 70], employees [26, 43, 49], or university faculty or students [6, 11, 14, 15, 31, 53, 86, 87]. Our target users were employees of a governmental organisation.

Forgetting Rate of Different Age Groups: [44] evaluated the ability to recall visual cues after 20-30 minutes and 75

¹Interventions can be tools or security awareness and education measures.

days. They did not find any age differences in the recall ability of these visual cues. [101] examined the recall ability of verbal cues after 1 and 62 days with different age groups. They conclude that the encoding of the information in the beginning is slower, but the rate of forgetting is comparable afterwards. *Retention Periods of IT-Related Training:* while most of the previously mentioned phishing studies evaluated the impact of their interventions straight after roll-out, a few also evaluated the effect after some time had elapsed. These mainly showed that the effect still held and did not systematically determine for how long the effect was still evident. These retention studies were mainly conducted in the context of security awareness and education measures. In [28, 60], retention was evaluated after approximately a month. In [116], retention was evaluated after 45 days. In [107], the retention was evaluated after 8 weeks. In [22], retention was evaluated after 5 months. All showed that the effect was still perceptible but was often no longer significant. [78] examined the ability to judge insecure password-related behaviour. The participants received awareness-raising materials and were tested again after 6 months. The participants were able to retain significant knowledge. *In our case, we study exactly how long the effect lasts.*

3 Use Case: Organisation Description

A State Office for Geoinformation and State Survey (SOGSS) is a public administration sector organisation. Its core activities relate to land register and real estate cadastre. Overall, SOGSS has about 2200 employees, 83% of whom have a technical background in either surveying, geodesy, geoinformatics or other related fields, such as photogrammetry. 60% of the employees are over 50 years of age. Only 14% are between 25 and 35 years of age. 40% of all employees are female, 60% are male. All employees use passwords and their SOGSS email account on a daily basis. Email communication with colleagues, citizens, and partners from business, science and other authorities is indispensable to employees.

SOGSS has nine regional head offices, a central operational office and a central head office, each with several departments and each in a different city. Like all organisations in the German public administration sector, SOGSS is required by the national IT Planning Council [71] to implement an information security management system (ISMS). Thus, SOGSS established the position of a chief information security manager (CISM) and the role of '*person of contact for information security concerns*' (PoC-InfoSec) was introduced as organisational interface between the CISM and the local offices. The managers of all ten human resources and administration departments (from nine regional head offices and the central operational office), as well as the manager of the central head office, perform this role. Furthermore, SOGSS decided to develop a security awareness & education programme containing one mandatory on-site tutorial, which was delivered

to all employees. Most tutorial sessions were held in October 2018.

4 Security Awareness and Education Measure

We describe the design decisions made for the mandatory measure rolled out in 2018. Afterwards, we describe their structure and content. Finally, we introduce the reminder measures.

4.1 Design Decisions

The organisation decided to use on-site tutorials instead of other delivery measures, such as web-based training, for two reasons. Firstly, on-site tutorials are common practice at SOGSS and therefore employees' acceptance of such tutorials was expected to be higher than for other formats. Secondly, the search for a suitable third-party web-based training, and the obligatory call for tenders, would have taken too long. Due to room size constraints, it was decided to deliver training to forty participants in each tutorial. Furthermore, based on the experiences from other on-site tutorials, it was decided that the tutorial should last three to four hours. Thus, the content had to fit into this allotted time.

A decision was made to adopt a 'train the instructor' approach, instead of having the CISM delivering all the tutorials. The 'train the instructor' approach was chosen since it represents a resource-efficient way to deliver tutorials to a large number of employees over a reasonably short period of time. The eleven PoC-InfoSec were trained by the CISM. To support them, a Power Point presentation was developed in two versions: 1) an instructor version, supplemented with explanations and instructions on how to facilitate audience interaction, and 2) the actual presentation to be used during the tutorial.

4.2 Content Overview

It was decided that the tutorial would address the following three topics, as threat reports and the organisation's experiences identified these as the most relevant ones: *Topic-1:* General security awareness, *Topic-2:* Phishing, and *Topic-3:* Password best practice. While the first and third parts were developed by the CISM from scratch, the second part was an adaptation of the awareness and education measure reported by [84]. Correspondingly, the focus of our investigation was on Topic-2². Its content will be described in more detail in the following two subsections.

²It could be argued that we should only have addressed this one topic in the tutorial if only one was going to be evaluated. However, this would not have been sufficient to be compliant with the ISMS. We could indeed have conducted two different tutorials, but this would have been much less efficient and there might have still been a bias because the study ran over several months and therefore everyone would have had to participate in the second tutorial in the same period of time, too. Thus, because we wanted to be able to conduct our evaluation in the field, we had to accept this trade-off.

With respect to the content of the tutorials, *Topic 1* provided a general introduction to the information security topic and information about how security incidents have to be dealt with in SOGSS. The organisation's threat statistics and typical threat vectors such as email (being one of the most common in public administration) were introduced. Where suitable, examples of anonymised in-house IT-security incidents were provided to demonstrate vulnerabilities when using the Internet. The content of *Topic 3* made employees aware of the risks of using weak passwords and introduced methods of building strong passwords. This part ended with a short interactive quiz, where examples were used to assess the respective password strength.

4.3 Topic 2 Content: Phishing

We customised a security awareness and education measure, which was developed and evaluated³ by Neumann *et al.* [84]. This material is very well suited: it has been evaluated in an organisational context attesting its effectiveness and it is freely available in German. The original content was prepared for self study use, i.e., reading a pdf or integrated it into an e-learning platform. Thus, it needed to be customised. The content of the measure had two parts:

Part-A provided general information about phishing, including: (a) why everyone can potentially become a victim, (b) that phishers don't just use email messages but any type of message, (c) that there are various types of phishing messages (including those asking for sensitive information, those including dangerous links, and those with dangerous files attached), (d) what the potential consequences of falling for a phishing attack are, and (e) the recommendation to delete phishing messages and to search for further information when the person is uncertain.

Part-B: commenced by explaining that a number of plausibility checks should be carried out, including checking the language, the style, and the sender information. Afterwards, the focus was on phishing messages which look plausible, at first glance, but which actually contain potentially dangerous links and/or attached files. Legitimate messages might be used as template, with the sender address being spoofed, and the URL behind a link and/or the attachment being replaced. *First*, employees were shown how to check whether an embedded link was dangerous to click on. This covers several attack types that phishers use to trick people (for more information see Appendix B). Furthermore, it was explained that these tricks are combined by phishers and that the presence of https is an unreliable indicator of trustworthiness. *Second*, they were shown how to identify dangerous file(s) and told which tricks phishers apply to trick people, incl. using two file extensions or unknown extensions.

³Note that effectiveness was evaluated straight after rolling-out the measure.

Both parts contained several example messages. All examples were synthetic as we came up with our own brands. There were example messages to illustrate various aspects of each attack type, as well as example messages to practice what was learned. Various misconceptions (such as that people tend to classify long URLs as Phishing links) identified from the literature were addressed throughout the measure, too.

Customization. Before customization the measure was usable only for self-studies: it is full text PDF documents that contains full sentences. This was not suitable for on-site tutorials, thus it had to be adopted to be used in on-site tutorials with Power Point presentations. Furthermore, *Part-A* (b), i.e. the information that phishers may use various message media, was not addressed in SOGSS's version. The use of email on mobile devices and/or social media is severely restricted to a small number of SOGSS's employees. Hence the tutorial focused on desktop application emails, as they are the only relevant target for phishers at SOGSS.

All examples from the original measure were replaced by ones more related to the employees' daily work. For example, where possible, anonymised examples from reported phishing emails were used. The PoC-InfoSecs were asked to show one example after the other. For each example, they were required to ask the audience whether it was a phish or not and one person from the audience was to justify the answer. Afterwards, the PoC-InfoSecs were supposed to explain the correct answer and comment on the answers given by the audience. This approach was used to attract the audience's attention. Finally, a summary of the most important findings and recommendations to check for was included at the end of the phishing part of the Power Point presentation.

4.4 Reminder Measures

Once the initial effect of the security awareness and education programme on-site tutorial has worn off, a reminder measure should be distributed to remind users of the information in the programme. Due to the lack of research into these measures, the goals of this research project was first to identify appropriate ones (i.e. by evaluating several). Correspondingly the reminder measures described in this section are currently – unlike the on-site tutorials – not part of the security awareness and education programme. So far, only participants in the corresponding study groups saw the reminder measures.

To the best of our knowledge, this aspect has not been studied in terms of which kind of presentation is most appropriate in the information security field. Four different types of reminder measures were developed and then evaluated: a text measure, a video measure, a interactive examples measure, and a short text measure. To inform the development of the reminder measures, we wanted to satisfy two requirements. *Firstly*, the measure must stand for itself. Apart from the shown measure, no further references or information should

be necessary. In particular, no instructions or introductions from another person should be required. *Secondly*, the measure content should match the on-site tutorial, i.e. it should not contain any new content that does not represent previously-learned knowledge from the on-site tutorial. The text and interactive examples measures contain exactly the same content (part A+B of the on-site tutorial). The video also covers this same content, but presents it as a story in order to make it more appealing. The short text cuts down the content to only include part-B and minimised descriptions of the attacks and defence strategies. In detail, the four reminder measures are:

Text: this measure is depicted in Figure 6 in the Appendix. It is a text, in German, with six figures. These visualise explanations such as the structure of URLs and that the actual linked URL is displayed in a status bar or a tooltip.

Video measure: this measure presents the same content as the previous one but relies mostly on visual explanations and narration, instead of text. Figure 7 in the Appendix gives an example of the video measure.

Interactive examples: this measure uses an interactive presentation. The content is presented as two interactive examples of phishing emails (see Figure 8 in the Appendix). Each of the emails has multiple interactivity-points marked with red dots, which reveals information about the respective part of the email when hovered over. In order to finish this measure, the trainee has to click at least once on each area.

Short Text: this measure represents a text-based measure with curtailed content compared to the previous measures. It contains only Part-B, i.e. it focuses only on the recommendations for detecting phishing messages (see Figure 9 in the Appendix).

5 Methodology

We first introduce the research questions and the hypotheses. Then, we discuss our study design decisions. Afterwards, we provide details about the study, i.e. recruitment, group assignment, used email examples, and actual study procedure as well as ethical considerations.

5.1 Research Questions

We want to answer three research questions. The first one is: **How long does the effect of the on-site tutorial last, i.e. when should the gained awareness and knowledge be reminded?**

The following pre-condition needs to hold: the measure significantly strengthens participants' skills in distinguishing between phishing emails and legitimate emails, straight after the on-site tutorial. Therefore we phrase the following hypothesis for this pre-condition:

$H_{M_{pre}-M_{0M}}$:Participants have an enhanced skill in terms of distinguishing between phishing and legitimate

emails directly after the on-site tutorial, i.e. 0 months after it, as compared to before participating in the on-site tutorial.

In order to investigate the effectiveness⁴ of the on-site tutorial over time, we formulate the following hypothesis for the continued testing:

$H_{M_{pre}-M_{\Delta iM}}$:Participants have an enhanced skill in terms of distinguishing between phishing and legitimate emails after $\Delta t = 4 + 2i$ months, where $i \in \{0, 1, 2, 3, 4\}$, as compared to before participating in the on-site tutorial.

We decided to start the follow-up evaluations after four months due to results from related work in the phishing context [22, 60, 116] reporting significant effects from security awareness and education measures lasting from 45 days to up to five months. Therefore either 45 days or five months after the on-site tutorial should be chosen for the first follow-up evaluation. However, conducting the first follow-up evaluation after 5 months would have increased the likelihood that the effect of the on-site tutorial had decreased below a significant improvement. Therefore, we decided on a more conservative approach, i.e. to start the follow-up evaluations earlier. Starting too early would have required too many participant groups due to the between-subjects approach. Therefore, we commenced the follow-up evaluation after four months, since it represented the best trade-off and allowed a meaningful study design despite the limited overall number of participants available. Note that the scheduled maximum duration of the evaluation was set to twelve months due to legislative constraints⁵ of SOGSS.

The second research question is: **Which of the four reminder measures performs best – straight after its roll-out?**

The reminder measures were distributed as soon as $H_{M_{pre}-M_{\Delta iM}}$ no longer held, i.e. after six months⁶.

To study this second research question, the following pre-condition needs to hold: the potentially best reminder measure needs to significantly strengthen participants' skills in distinguishing between phishing emails and legitimate emails - right after the reminder measure was rolled out. Correspondingly, we use the following hypothesis:

⁴When we talk about effectiveness from now on, we only talk about the effectiveness of the phishing part of the on-site tutorial.

⁵The next security awareness and education measure is required after one year due to compliance reasons

⁶We analysed the data as soon as they arrived to distribute the reminder measures as soon as the performance is no longer significantly better as compared to before they participated in the on-site tutorial. As described in Section D.1, our evaluation yields the result $\Delta\bar{t}=6$, i.e. the measurement after six months did not detect a significantly enhanced skill in terms of distinguishing between phishing and legitimate emails. We provide this information here, to facilitate description of the remaining two research questions.

$H_{M_{pre}-M_{Reminder_x-6M}}$: Participants have an enhanced skill to distinguish between phishing and legitimate emails directly after the distribution of reminder measure $Reminder_x \in \{\text{Text, Video, Interactive examples, Short Text}\}$, as compared to before participating in the on-site tutorial.

For those reminder measures for which this pre-condition holds, we compare the measured effects, to see one is more superior than any others.

The third research question is: **How long does the effect of reminder measures last?** To measure this, we evaluated the performances of those reminder measures for which the precondition from the second research question holds in a six month retention (i.e. 12 months after the roll-out of the on-site tutorial). The corresponding hypothesis is:

$H_{M_{pre}-M_{Reminder_y-12M}}$: Participants still have an enhanced skill to distinguish between phishing and legitimate emails six months after distribution of the measure for which $H_{M_{pre}-M_{Reminder_x-6M}}$ holds, compared to before participating in the on-site tutorial.

This research question has two pre-conditions: (1) participating twice (once after six months and again after twelve months) should not have a significant impact on the measured effect (2) other events in the organisation should not lead to a significant improvement again.

5.2 Design Decisions for Study Design

The selection of the study type was driven by the need to gain a high participation rate and outcome quality of the study within the SOGSS environment. We decided on a study design that would enable remote participation. This allowed us to reach more participants than a lab study. Moreover, remote participation was less time consuming for the SOGSS' employees (being distributed over several locations and cities) and it was less likely to interrupt their work as they could participate during the allotted time frame at their convenience.

There are two main ways of conducting the evaluation of the on-site tutorial with remote participation: a multiple choice test, e.g. asking to define phishing and name attack types; or evaluating participants' actual skill to identify phishing and legitimate emails. Multiple choice tests would have provided very little information about the enhancement of employees' skills in terms of distinguishing phishing and legitimate emails from each other as we would not be able to determine whether emails' other properties (e.g. the deceptively trustworthy design or the sender name) may have led them to judge phishing emails as legitimate (or the other way round). We decided to employ the second option which asked participants to classify emails as either phishing or

legitimate.

There are two main ways to evaluate participants' actual skills in terms of distinguishing between phishing and legitimate emails: (i) sending them phishing emails (with or without announcing the fact that phishing emails will be sent) and then e.g. asking them to report phishing emails; or (ii) displaying a set of emails in a survey style and asking them to decide which were phish and which were legitimate. The first option might be considered to be closer to assessing real behaviour but such behaviour might well be influenced by the fact that participants know they will receive phishing emails, but they don't know when they will receive them. Their daily business remains their main task, not security, which is ecologically valid. In contrast, the second approach measures skills in a 'best case' scenario as security is the participants' primary goal in this case, and this is unrealistic.

However, the first approach – actually sending phishing emails – was not feasible in our study setting at SOGSS. Some reasons were: the research goal of assessing gained awareness and knowledge over time required us to evaluate all data in a very short time span for each group (i.e., for each time Δt_i). To ensure that the received phishing emails amount was realistic, we could not have sent more than one test phishing email per day – but also not send one every day. As the goal was to evaluate skills for all five attack types, this evaluation have taken too long. Moreover, sending phishing emails to employees of German organisations would have required extra permissions e.g. from work councils. There are also some general issues with this approach regarding data collection quality (as e.g. discussed by [83]). For these reasons, we employed the first approach – displaying email screenshots in an online survey, where participants assessed all emails in one session and decided, for each, whether it was a phish or legitimate. This allowed us to evaluate all attack types in the shortest possible amount of time.

5.3 Recruitment and Group Assignment

Due to SOGSS organisational requirements, participation in the on-site tutorial was mandatory, but participating in the evaluation was voluntary. The information about the evaluation and a corresponding link to the survey was emailed to employees by SOGSS's CISM. Every group only got one survey sent to them (we did not reuse groups/participants for other groups). Once they received the notification to take part in the survey, they had a week to do so. A reminder email was sent that emphasised the importance of personal participation due to the cyber security situation. This was sent to everyone in the group as it was not known who had actually participated as yet. We collected data for two weeks.

We planned for eleven groups (see Figure 1): seven retention groups and four reminder groups. For the assignment, we considered the fact that besides the October on-site tutorials, a few were scheduled for end of 2018 / beginning of

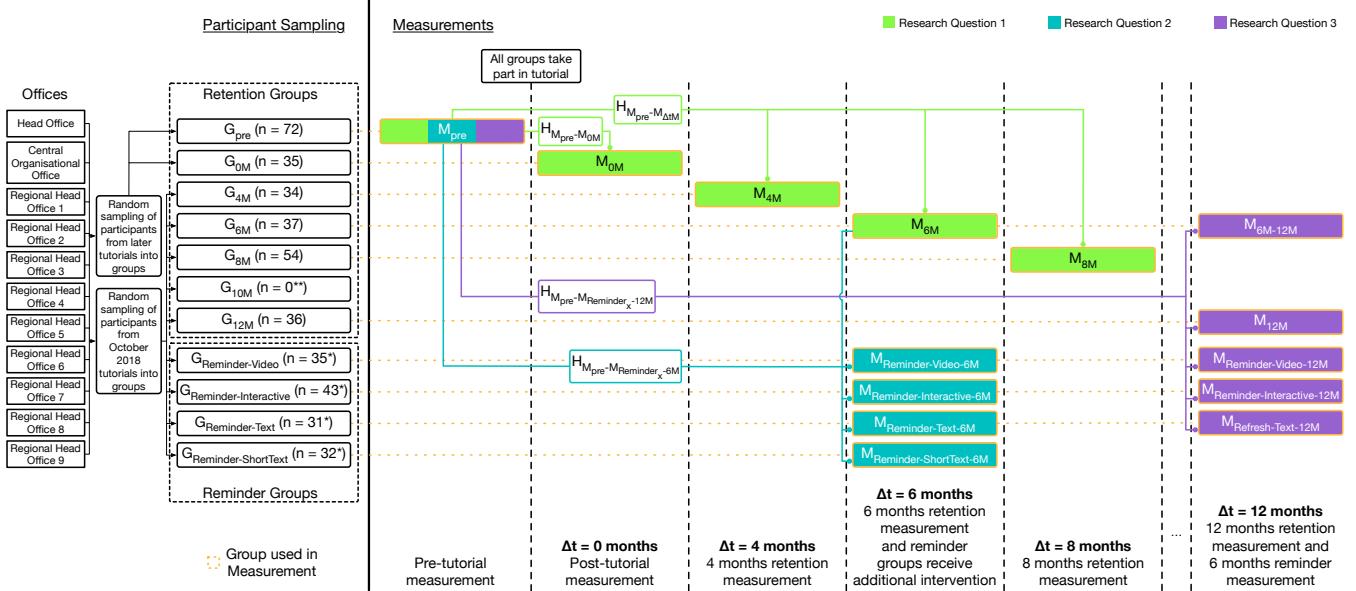


Figure 1: Overview of the participant sampling into the ten study groups, the measurements taken for each of the groups, and the hypotheses in relation to the measurements used in the corresponding statistical tests. Group sizes marked with an ‘*’ are the overall group size, not those participants used for the linked analyses (see Section 6.2). Group G_{10M} (marked with ‘**’) was originally scheduled and participants assigned to it, but due to the lack of significant effects in G_{6M} and G_{8M} the group was never tested. The • indicates for each hypothesis in which measurement the participants are expected to perform better.

2019 to enable those employees who could not attend any of the October options to participate. To prevent the introduction of variance into the later measurements, all participants from these later tutorials were randomly assigned to either G_{pre} or G_{0M} . All other participants were randomly assigned to one of the retention or reminder groups. Thereby, we ensured that participants from the same office were equally distributed among all groups. Thus, participants from each office were represented in each group. This was important as employees were taught by different instructors at different locations. Since it was a requirement of SOGSS, we did not collect any demographic data in the study, and no other parameters were used for the sampling. Every measurement consisted of unique individuals except for the reminder measurements at month 12 and the measurement at 12 months for measuring participants twice. The numbers in Figure 1 reflect these unique participants. The linked participants are a part of the full reminder measurements that we could link based on a code they entered for the 6-month measurement and the 12-month measurement. We will add this description to section 5 to make the distinction clearer and earlier. For $G_{Reminder-Interactive}$, $G_{Reminder-Text}$, $G_{Reminder-Video}$, and G_{6M} , the measurement after 12 months is longitudinal. Therefore, a subset of the overall participants from e.g. $G_{Reminder-Interactive}$ after 6 months build the corresponding group after 12 months. We denote those participants in each of the four

subsets as linked. The overall participants in each group after 6 months are called unmatched.

Note that all assignments were implemented by the CISM for data protection reasons.

5.4 Email Screenshots

As outlined in Section 5.2, the participants’ performance was measured using screenshots of emails, each of which has to be classified as phish or legitimate. In an ideal setting, one would have evaluated all combinations, relevant to SOGSS, of five attack types (see Section 4.3), operating systems (including at least Windows and macOS), email clients (including at least Outlook, Thunderbird, Apple Mail), and web browsers (including at least Firefox, Chrome, Safari). However, this would have resulted in an infeasibly large number⁷ of screenshots that each participant would have needed to rate, leading to fatigue effects or abandonment. Therefore, we selected a representative subset of all possible screenshots. The goal was to cover all attack types with a variability within the remaining characteristics. (This resulted in ten different phishing email screenshots. This set of phishing email screenshots was complemented by an equal set of ten legitimate email screen-

⁷ At least $180 = 5 \times 2 \times 3 \times 3 \times 2$, assuming there are as many phishes as legitimate emails.

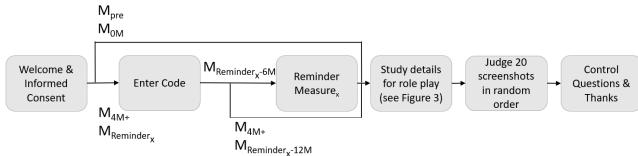


Figure 2: The survey process and measurement times.

shots⁸. We decided to use ‘https’ for all URLs (both phishing and legitimate)⁹. The screenshots with a dangerous link were generated in a way that the mouse was already situated next to this link, i.e. the actual URL was displayed depending on the environment in the tooltip or the status bar. The screenshots of the phishing emails used in our evaluation are provided in the supplementary material (see Figure 11 to Figure 15 in the appendix). The following phishing emails were used: one easily detected phishing email with implausible email content: a dubious job offer, or an offer of unrealistic amounts of money. Eight phishing emails with plausible content, but including dangerous links: four emails presenting the URL in the toolbar and four in the status bar; two with a mismatch and one with a faked tooltip; URLs were either arbitrary, had the domain name as subdomain, a typo in the domain, or the domain name was extended. One phishing email with plausible email content, including a dangerous attachment.

5.5 Survey Design

The survey was designed to match the corporate design of the organisation and it was implemented on the SoSci Survey platform¹⁰. The overall structure is depicted in Figure 2.

For all measurements, after opening the corresponding SoSci Survey link, participants received *explanations* about the evaluation and the cooperation between their employer (SOGSS) and our university for this evaluation. We tried to mitigate external factors by also explaining that participants ought not to use external sources (e.g. web search) and to concentrate on the survey during the session. We tried to increase participation by highlighting the possibility of using this evaluation as a self-assessment of their own skills. Moreover, we assured the participants that they did not need to fear consequences if they performed poorly. Thereafter, the survey’s structure was explained to them. Participants were informed that they could terminate their participation at any time without providing any reason and that, in this case, their

⁸While this does not represent the usual ratio of phishing to legitimate emails employees might find in their inboxes, it is appropriate for knowledge assessments in particular to compare the performance against guessing

⁹The alternatives were either to use only http (which could have led to a misleading message that dangerous links always use http and a conclusion that https is always secure) or to combine each attack type applied with both which would have lead to twice as many email screenshots.

¹⁰<https://www.soscisurvey.de/>

data would not be used (*informed consent*). Next, participants from all groups except G_{pre} and G_{0M} saw a page where they were supposed to enter an individual self-selected *code*. This code was necessary to permit us to link initial and subsequent measurements (without violating anonymity).

Afterwards for the $M_{Reminder-x-6M}$ the corresponding *reminder measure* was displayed with a short introduction of what to expect on the page. The three measures were designed to be similar: text, video and interactive examples required the same amount of time (8 minutes). Short Text, which was intentionally shorter, required 3 minutes to complete. Due to a technical error interactive examples was also set to 3 minutes.

Next, all groups saw the same page again. We used a *role play* approach. We told participants, before displaying the screenshots, that they should assume that they are someone called ‘Martin Müller’. Relevant details about Martin were provided (see Appendix C). Then, the *email screenshots were shown in a random order*, one per page. For each of the 20 screenshots, they had to decide whether it was a phish or legitimate. At the end, participants had to answer a few *control questions* such as their usage of the Internet or revising the tutorial material. For the measurements after twelve months, participants were also asked whether they had already participated after six months.

5.6 Ethics and Data Protection

Participation in the evaluation was optional and the survey could have been completed at a time of the participant’s choice. Participation was not remunerated in any way but they could have participated during their working hours. Due to strong privacy regulations in Germany, the anonymity of participants was a mandatory requirement. Therefore, we used SoSciSurvey to collect the data (they are compliant with the new European Data Protection Regulation). The previously described process to assign participants to groups, to invite and to remind them, as well as the fact that no demographic data was collected, was discussed with and approved by the works council, as this prevented any kind of individual performance monitoring. All information about the process, the anonymity, the agreement of the works council, as well as the fact that they did not need to fear consequences for poor performance was provided to the participants in the invitation email. It was also advised to get in touch with the CISM in case of any ambiguity or questions about the received email.

6 Results

We first provide information about our participants and then present the results for each of our three research questions.

6.1 Participants and Data Cleaning

A total of 439 participants completed the online survey (several due to the two measure points for research question 2 and 3). We performed the following data cleansing steps: (1) We excluded four participants whose answers evidenced specific patterns. They had 100% phishing email identification and 0% legitimate email identification respectively; i.e., they judged all emails as phishing emails. (2) We excluded 26 participants who admitted using the Internet or other sources to answer or right before answering the questions. Thus, the data from 409 participants was analysed.

6.2 Analysis Methods

We used the Signal Detection Theory (SDT) [97] to measure the participants' performance, i.e. whether participants' skills in distinguishing between phishing and legitimate emails improved, as compared to before the tutorial. This theory has been used in other studies evaluating phishing identification [14–19, 38, 75, 76, 79, 82, 93, 94]. SDT enables us to discern between signal (phishing emails) and noise (legitimate emails). In line with above-mentioned literature, we used the following two output values: sensitivity (d') and criterion (C). In the context of our research, sensitivity defined the skill to distinguish phishing emails (signal) from legitimate ones (noise). The larger d' , the better the participants' performance in distinguishing signal from noise. Criterion (C) was defined as the response tendency, e.g. in our case whether participants were more cautious after the tutorial, i.e. more legitimate emails were classified as phishing (more false negatives), or did they take more risks, i.e. more phishing emails are classified as legitimate (more false positives). The closer this criterion was to 0, the more accurately they decided whether a signal was phish or legitimate.

We evaluated the assumptions relevant for calculating SDT parameters, i.e. equal variance and Gaussian distribution. Afterwards, we calculated the SDT parameters for sensitivity and criterion per participant. We then calculated the mean values per measurement using SPSS. To evaluate our hypotheses, we analysed the differences for participants' sensitivity and criterion values using one-way ANOVAs (using SPSS). ANOVA is a common tool to analyse forgetting curves as it overcomes the problem of initial learning levels [44]. For every ANOVA, we started off the analysis by checking the assumptions for both the sensitivity and the criterion. Since both sensitivity and criterion only violated the normal distribution assumption, and the ANOVA is relatively robust against the violation of this assumptions [42], we continued the analysis. For the descriptive results for the sensitivity see Figure 3, 4, and 5. The hit-rates for phishes and legitimate results are provided there as well. For simplicity and readability reasons we will only state the significant results in the following subsections (for full version see Appendix D).

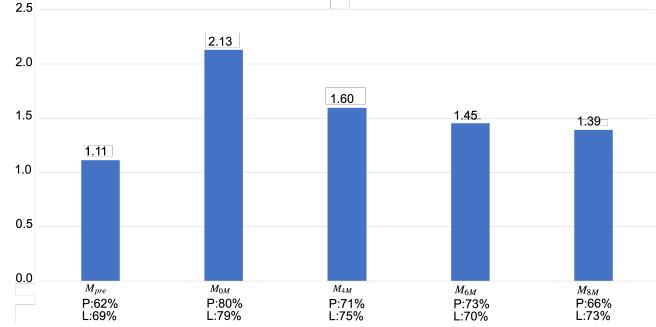


Figure 3: Sensitivity score and hit rates (Phish/Legitimate) for the measure of RQ1.

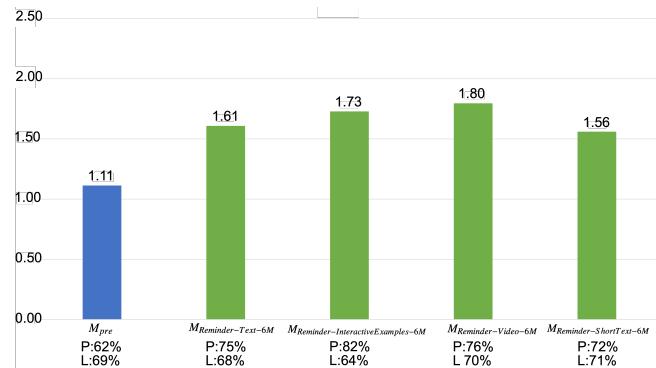


Figure 4: Sensitivity score and hit rates (Phish/Legitimate) for the measure of RQ2 (blue = only tutorial, green = reminder measure groups).

6.3 Results for Research Question 1

As stated before, we analysed the data as soon as possible so that we could distribute the reminder measure as soon as the performance was no longer significantly better as compared to their performance before they participated in the on-site tutorial. We discovered that after a period of six months, performance was no longer significantly different from before the on-site tutorial. However, we decided to continue collecting data after eight months to strengthen our findings. We wanted to make sure that the difference between those two groups was not due to variance of the participants in our between-subjects design.

For the reporting of the results, we combine the analyses of $H_{M_{pre}-M_0M}$ and $H_{M_{pre}-M_{\Delta t M}}$ (for $\Delta t = 4 + 2i$ months, where $i \in \{0, 1, 2\}$) We checked for a significant difference between the corresponding five groups (see Figure 1) using a one-way ANOVA. There was statistical significance between the groups ($F(4, 227) = 5.457, p < 0.001$) for the sensitivity (d'). For the effect size we calculated $\omega^2 = .093$, which is a medium effect size according to [42]. A LSD post-hoc showed that

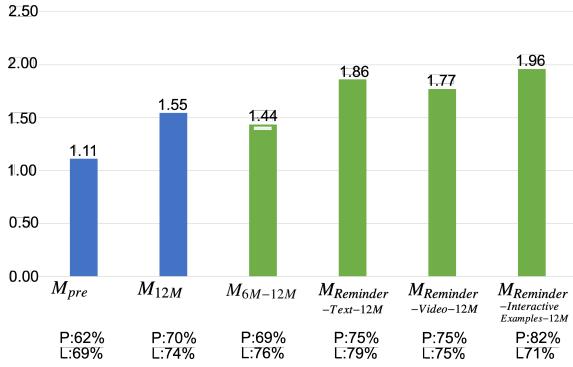


Figure 5: Sensitivity score and hit rates (Phish/Legitimate) for the measure of RQ3 (blue = only tutorial, green = reminder measure groups).

the sensitivity for the M_{0M} ($d' = 2.13$, $SD = 1.15$) was significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$) with ($p < .001$). The LSD post-hoc test showed that the sensitivity for the M_{4M} ($d' = 1.60$, $SD = 1.01$) was significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$) with ($p = .034$). Note, there was no statistical significance between the groups ($p = 0.623$) for the criterion (C).

In summary: We accept $H_{M_{pre}-0Month}$ and $H_{M_{pre}-4Months}$.

6.4 Results for Research Question 2

First, we checked for which reminder measures the hypothesis $H_{M_{pre}-M_{Reminderx-6M}}$ holds. We checked for a significant difference between M_{pre} and the four months retention groups (see Figure 1) using a one-way ANOVA. There was statistical significance between the groups ($F(5, 244) = 2.410$, $p = 0.037$) for the sensitivity (d'). For the effect size we calculated $\omega^2 = .027$, which is a small effect size [42]. A LSD post-hoc showed that the sensitivity for the $M_{Reminder-Text-6M}$ ($d' = 1.61$, $SD = 1.18$) with ($p = .005$), $M_{Reminder-Video-6M}$ ($d' = 1.80$, $SD = 1.42$) with ($p = .005$) and $M_{Reminder-InteractiveExamples-6M}$ ($d' = 1.73$, $SD = 1.19$) with ($p = .007$) were significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$). Note, there was statistical significance between the groups ($p = 0.013$) for the criterion (C). A LSD post-hoc showed that the criterion for the $M_{Reminder-Text-6M}$ ($C = -.23$, $SD = .59$) with ($p = .043$) and $M_{Reminder-InteractiveExamples-6M}$ ($C = -.43$, $SD = .65$) with ($p < .001$) were significantly different from the M_{pre} ($C = .12$, $SD = .84$).

In summary: We accept $H_{M_{pre}-M_{Reminderx-6M}}$ for text measure, video measure, and interactive examples measure.

In order to test whether one of the three remaining reminder measures performs best, we also checked the ANOVA values for between the reminder measures. There is no significant difference between these measures. From the descriptive data,

the interactive examples measure performs slightly better than the video measure (see Figure 5).

6.5 Results for Research Question 3

Based on the results from RQ2 we decided to not collect data from the short text group after 12 months. In order to address the pre-conditions from Section 5.1 we kind of extended $H_{M_{pre}-M_{Remindery-12M}}$ accordingly, i.e. six measurements were considered (see Figure 2).

We linked participants using the provided codes. This resulted in 20 participants in M_{6M-12M} , 17 participants in $M_{Reminder-Text-12M}$, 17 participants in $M_{Reminder-Video-12M}$, and 12 participants in $M_{Reminder-InteractiveExamples-12M}$. We analysed the data from participants that we could link via code. We checked for a significant difference between the corresponding six measurements. There was statistical significance between the groups ($F(5, 172) = 2.721$, $p = 0.022$) for the sensitivity (d'). The LSD post-hoc showed that the sensitivity for the $M_{Reminder-Text-12M}$ ($d' = 1.93$, $SD = 1.17$) with ($p = .009$), the $M_{Reminder-Video-12M}$ ($d' = 1.77$, $SD = 1.32$) with ($p = .031$) and $M_{Reminder-InteractiveExamples-12M}$ ($d' = 1.96$, $SD = 1.34$) with ($p = .016$) were significantly higher than for the M_{pre} ($d' = 1.11$, $SD = 1.12$). For the effect size we calculated $\omega^2 = .047$, which is a small effect size according to [42]. Note, there was no statistical significance between the measurements ($p = 0.274$) for the criterion (C).

In summary: The pre-conditions hold and we accept $H_{M_{pre}-M_{Remindery-12M}}$ for video measure and interactive examples measure.

7 Discussion

We first discuss some general implications of our study, then our results for the three research questions and then the limitations of our work.

We excluded four participants because they marked all screenshots in the survey as phishing emails and therefore had 100% phishing detection but also 0% legitimate detection. In addition, we excluded 26 participants for seeking help for answering the survey. Seeking help is very useful in the real world. But as we could not control what kind of help they got and we explicitly mentioned that they should fill out the survey without help, we decided to exclude those that violated our rule.

We discussed advantages and disadvantages of publishing the results with the organisation. In particular the potential risk to the organization caused by publishing the results was evaluated. Together with the organisation it was decided to name a few key facts about the organization. We wanted to give the opportunity to other researchers to know the study setting in order to allow transferring the information to other contexts and making sure that our results can be correctly interpreted.

7.1 Discussion of RQ-1

The participants' skills in identifying phishing messages improved significantly straight after attending the tutorials. Our results are in line with those of the study evaluating the effectiveness of the original measure [84]. Thus, the customisation, as well as the switch to an instructor-based tutorial approach, seems not to have affected the efficacy of the content.

While the pre-condition for the first research question holds and the phishing detection rate increased from 62% to 80%, a closer inspection of the results begs the question whether an 80% phishing identification rate in the M_{0M} measurement leaves the participants sufficiently protected, considering the fact that security was their primary goal in our evaluation. After some internal deliberation about these numbers, the improvements were seen as a success at SOGGS, since this was the first organisation-wide security awareness and education measure and employees' skills in distinguishing legitimate and phishing emails were significantly enhanced by the tutorial. We will again deliver security awareness and education measures at SOGGS with the goal of further increasing these numbers.

With respect to the performance over time, we found that after six and eight months participants' skills were no longer significantly better than before participating in the tutorial. This also aligns with prior research reporting results of retention tests, albeit all of these studies had shorter retention periods. While they all used different interventions and also different evaluation techniques, they all found that the effect lasted until they conducted the retention study (which was max. after 5 months in [22]).

Our results show that current reminder periods required by standards such as PCI-DSS [27] – which usually require an interval of twelve months – should be re-considered.

We are aware that the SOGGS has a higher age average, as 60% of the employees are older than 50 years. One might assume that we would have achieved different results with younger participants. According to [44, 101] the age does not increase the forgetting rate significantly. Therefore, we argue that our results also hold true for younger people.

7.2 Discussion of RQ-2

Concerning the results for the reminder measures, both the video measure and the interactive examples measure stand out from the others in terms of sensitivity (1.797 and 1.728 versus 1.559 for the text measures after 6 months; and the text measure not being significantly better after 12 months compared to the measurement before the tutorial). Thus, our results indicate that – in line with related work such as [25, 67] – static measures lead to a poorer performance than dynamic measures. Furthermore, our results show that even short reminder measures can be very effective and it is neither necessary nor recommended (because of the time needed) to use the main

security awareness and education measure as reminder measure. Yet, it must also be noted that there is a lower bound to the information which must be included in the reminder material, as evidenced by the insufficient performance of the short text measure.

Overall, for SOGGS, the combination of costly on-site tutorials and an efficient reminder measure after six months looks very promising.

Whether the video measure or the interactive examples measure perform better is not that easily answered as there is no significant difference between the two. Considering the criterion values, we could argue the video measure achieved the best results. With respect to sensitivity, there is no clear 'winner' after 12 months. Note that although the performance after 6 months for the video measure and interactive examples measure is not significantly lower than directly after the on-site tutorial, the ultimate goal must be to get as close as possible to the performance achieved directly after the tutorial (see Figure 4).

7.3 Discussion of RQ-3

Following the discussion of RQ-1, the performance of participants who received the video measure and the interactive examples measure is so good that a refreshment might not even be necessary after another six months has elapsed. Such results should also be taken into account when reconsidering time intervals of international standards, as discussed in Section 7.1. In order to know when subsequent reminders should be scheduled, future research into their long-term effects is required.

7.4 Recommendations for Future Studies

We faced several challenges during our research. We discuss here how they were addressed to assist researchers planning similar studies - which we would welcome.

One challenge with our study was to avoid reporting an effect for our reminder measures while the effect was actual caused by external factors: e. g. media reports, or internal discussions. This was addressed by including the 12 month group and comparing their performance to those of the reminder groups.

The next challenge is that the exact point in time when we measured the performance could not be controlled (also in comparison to the participation in the tutorials which were offered for an entire month). To address this, we limited the time span for filling out the survey to two weeks after having received the invitation email. Note that while this might be an unusual design choice, this is reasonable in a field study as not all employees would receive a security awareness and education measure on the same day.

7.5 Limitations

Even though all instructors used the same measures and the same instructions for conducting the tutorial, it is still possible that there were small differences in the course of the held tutorials. Some of these groups were trained by the CISM herself. *This limitation was mitigated by the random assignment of employees from different locations to study groups.*

Our study design selected measures for best case scenarios, with security being the primary goal. We argue that it is worth testing in such scenarios as it is a pre-condition for identifying a phishing email during any working day. The results show that this pre-condition is far from being a given (the M_{pre} only detects 63% of the phishing emails). Furthermore, most of the example emails could only be identified as such when the URL behind the link and/or the file type were checked. Thus the phishing emails used in the evaluation were more difficult to identify as compared to the average phishing emails received by SOGSS employees. *For the purpose of our research, best case and poor performance before participating in the tutorial, this approach is appropriate.* We acknowledge that for statements in the actual working environment with actual received phishing emails, the study design would need to be different.

In addition, due to the restrictions of SoSci Survey, we could only provide screenshots, i.e. it was not necessary for participants to move the mouse to the link as it was already in the correct position, with the URL displayed. Furthermore, it was not possible to check several of the integrated links or get additional information such as the html source of the email. Thus, on the one hand we made it a bit easier and on the other hand a bit more difficult as compared to actual phishing email detection. Thus, in reality, the detection rate of the evaluated phishing emails is likely to be different if employees would have received them in their inboxes and were asked to judge them there. However, this would have made a lab setting necessary, which was not possible.

In order to keep the duration of the study feasible, we were restricted in the number of evaluated phishing emails. We selected a representative sample of emails from the large variety of possible options. It might be that different combinations would have given us different results. However, we believe that due to the selection of representative examples, the findings would not have been significantly different.

Our study was customised for, and conducted at, a German public sector organisation (SOGSS). Therefore, our participant sample is biased by the type of work and the technical background of the employees. We would need to replicate our study in other types of organisations and organisations with different employee characteristics. This is part of future work.

Due to a technical error, the group interactive examples had to spend only 3 minutes engaging with the material and not 8 minutes as planned (similar to text and video). Even though this meant that the participants spent less time with

the interactive examples measure than initially planned (about 7.3 minutes on average see Table 5 in the appendix), it still produced excellent results. There is currently no evidence to suggest that a more extended time spent with the material would have had a negative effect. Since, despite the aforementioned technical error, the measure achieved significant results with both $M_{Reminder-InteractiveExamples-6M}$ and $M_{Reminder-InteractiveExamples-12M}$, one would expect either the same or a better effect over a more extended period. *Therefore, we believe that this did not impact our results.*

Finally, it is worth mentioning that our results for phishing hold although in the tutorial three topics were addressed (and not just phishing). Thus, it might be that the effect would last slightly longer if only one topic were addressed. It is open to discussion which scenario is more realistic (a single topic tutorial or one with some similar topics).

8 Conclusion

We presented a study on how effective security awareness and education measures are over time, and what the best way is to remind users' awareness and knowledge. To this end, we carried out a field investigation within a German "State Office for Geoinformation and State Survey". We considered three research questions: i) *How long does the effect of the on-site tutorial last, i.e., when should the gained awareness and knowledge be reminded?*, ii) *Which of the four developed reminder measures performed best?*, and iii) *How long does the effect of reminder measures last?*.

From the almost 2000 employees, 409 voluntarily participated. From the fourth month after the on-site tutorial, we evaluated groups every two months to measure awareness and knowledge retention. After six months, we saw no improved performance in distinguishing phishing and legitimate emails. Four reminder measures were distributed to four groups (one per group): a) text, b) video measure, c) interactive examples, and d) a short text. Twelve months after the tutorial, we compared the knowledge retention of the four reminder groups with the the pre-group. Among the four reminder measures, the video measure and the interactive examples measure performed best, with their impact lasting at least six months after being rolled-out.

Acknowledgments

The research reported in this paper has furthermore been supported by the German Federal Ministry of Education and Research within the framework of the project KASTEL_SKI in the Competence Center for Applied Security Technology (KASTEL).

References

- [1] Pieter Agten, Wouter Joosen, Frank Piessens, and Nick Nikiforakis. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In *Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015)*. Internet Society, 2015.
- [2] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *22th USENIX Security CHI*, 2013.
- [3] Abdullah Alnajim and Malcolm Munro. An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection . In *6th International Conference on Information Technology: New Generations*, pages 405–410. IEEE, 2009.
- [4] Abdullah Alnajim and Malcolm Munro. An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection. *2009 Sixth International Conference on Information Technology: New Generations*, pages 405–410, 2009.
- [5] Abdullah Alnajim and Malcolm Munro. An approach to the implementation of the anti-phishing tool for phishing websites detection. In *2009 International Conference on Intelligent Networking and Collaborative Systems*, pages 105–112. IEEE, 2009.
- [6] Ibrahim Alseadoon. *The impact of users' characteristics on their ability to detect phishing emails*. PhD thesis, (Doctoral dissertation, Queensland University of Technology), 2014.
- [7] Mohamed Alsharnoubi, Furkan Alaca, and Sonia Chisson. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, 82, 2015.
- [8] Kholoud Althobaiti, Ghaidaa Rummani, and Kami Vaniea. A review of human-and computer-facing url phishing features. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 182–191. IEEE, 2019.
- [9] Kholoud Althobaiti, Kami Vaniea, and Serena Zheng. Faheem: Explaining urls to people using a slack bot. In *Symposium on Digital Behaviour Intervention for Cyber Security (AISB)*, pages 1–8, 2018.
- [10] Nalin AG Arachchilage, Ivan Flechais, and Konstantin Beznosov. A game storyboard design for avoiding phishing attacks. In *Proceedings of the 11th Symposium On Usable Privacy and Security (SOUPS)*, 2014.
- [11] Nalin Asanka Gamagedara Arachchilage, Steve Love, and Konstantin Beznosov. Phishing threat avoidance behaviour: An empirical investigation. *Computers in Human Behavior*, 60:185–197, 2016.
- [12] Zinaida Benenson, Freya Gassmann, and Robert Landwirth. Unpacking spear phishing susceptibility. In *Financial Cryptography and Data Security*, 2017.
- [13] Clemens Bergmann and Gamze Canova. Design, implementation and evaluation of an anti-phishing education app. Master's thesis, Technische Universität Darmstadt, 2014.
- [14] M Butavicius, K Parsons, M Pattinson, A McCormac, D Calic, and M Lillie. Understanding Susceptibility to Phishing Emails: Assessing the Impact of Individual Differences and Culture. *International Symposium on Human Aspects of Information Security & Assurance (HAISA)*, 2017.
- [15] Marcus Butavicius, Kathryn Parsons, Malcolm Pattinson, and Agata McCormac. Breaching the human firewall: Social engineering in phishing and spear-phishing emails. *Australasian Conference on Information Systems*, 2016.
- [16] Casey Canfield, Alex Davis, Baruch Fischhoff, and Forget -A on Usable Replication: Challenges in using data logs to validate phishing detection ability metrics. *Symposium on Usable Privacy and Security (SOUPS)*, 2017.
- [17] Casey Canfield, Baruch Fischhoff, and Alex Davis. Using Signal Detection Theory to Measure Phishing Detection Ability and Behavior. In *SOUPS*, 2015.
- [18] Casey Canfield, Baruch Fischhoff, and Alex Davis. Quantifying Phishing Susceptibility for Detection and Behavior Decisions. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 58:1158–1172, 2016.
- [19] Casey Inez Canfield and Baruch Fischhoff. Setting Priorities in Behavioral Interventions: An Application to Reducing Phishing Risk. *Risk Analysis*, 38:826–838, 2018.
- [20] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Roland Borza. NoPhish: An Anti-Phishing Education App. In *Security and Trust Management*, pages 188–192. LNCS, 2014.
- [21] Gamze Canova, Melanie Volkamer, Clemens Bergmann, Roland Borza, Benjamin Reinheimer, Simon Stockhardt, and Ralf Tenberg. Learn to Spot Phishing URLs with the Android NoPhish App. In *WISE 9*, pages 87–100. Springer, 2015.

- [22] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Benjamin Reinheimer. NoPhish App Evaluation: Lab and Retention Study. In *USEC*. Internet Society, 2015.
- [23] Madhusudhanan Chandrasekaran, Krishnan Narayanan, and Shambhu Upadhyaya. Phishing E-mail Detection Based on Structural Properties. *NYS Cyber Security conference*, 2006.
- [24] Sidharth Chhabra, Anupama Aggarwal, Fabricio Benvenuto, and Ponnurangam Kumaraguru. Phi. sh/\$ocial: the phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, pages 92–101, 2011.
- [25] Hee Jun Choi and Scott D. Johnson. The Effect of Context-Based Video Instruction on Learning and Motivation in Online Courses. *American Journal of Distance Education*, 19(4):215–227, 2005.
- [26] Dan Conway, Ronnie Taib, Mitch Harris, Shlomo K Berkovsky, Kun Yu, and Fang Chen. A qualitative investigation of bank employee experiences of information security and phishing. *Symposium on Usable Privacy and Security (SOUPS)*, 2017.
- [27] PCI Security Standards Council. PCI DSSv3.2.1 - May 2018, 2018. https://www.pcisecuritystandards.org/document_library accessed: 27.02.2020.
- [28] Eugène JFM Custers. Long-term retention of basic science knowledge: a review study. *Advances in Health Sciences Education*, 15(1):109–128, 2010.
- [29] Rachna Dhamija and J. D. Tygar. The battle against phishing: Dynamic Security Skins. In *Symposium on Usable Privacy and Security*, pages 77–88, New York, NY, USA, 2005. ACM.
- [30] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why phishing works. In *Proceedings of CHI 2006 Human Factors in Computing Systems*, pages 581–590. ACM, 2006.
- [31] Alejandra Diaz, Alan T Sherman, and Anupam Joshi. Phishing in an academic community: A study of user susceptibility and behavior. *Cryptologia*, pages 1–15, 2019.
- [32] Matt Dixon, Nalin AG Arachchilage, and James Nicholson. Engaging Users with Educational Games: The Case of Phishing. *Woodstock*, 2019.
- [33] Ronald C Dodge, Curtis Carver, and Aaron J Ferguson. Phishing for user security awareness. *Computers and Security*, 26(1):73–80, Elsevier, 2007.
- [34] Julie S Downs, Mandy Holbrook, and Lorrie Faith Cranor. Behavioral response to phishing risk. *Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit*, pages 37–44, 2007.
- [35] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. Decision strategies and susceptibility to phishing. In *Proceedings of the Second Symposium on Usable Privacy and Security*, pages 79–90, 2006.
- [36] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *CHI*, pages 1065–1074. ACM, 2008.
- [37] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2873–2882, 2015.
- [38] Iain Embrey and Kim Kaivanto. Many Phish in the C: A Coexisting-Choice-Criteria Model of Security Behavior. *arxiv*, 2018.
- [39] Jussi-Pekka Erkkilä. Why we fall for phishing. In *Conference on Human Factors in Computer Systems*. ACM, 2011.
- [40] Adrienne Porter Felt, Robert W Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Embre Acer, Elisabeth Morant, and Sunny Consolvo. Rethinking connection security indicators. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 1–14, 2016.
- [41] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *WWW*, pages 649–656. ACM, 2007.
- [42] Andy Field. *Discovering statistics using IBM SPSS statistics*. sage, 2013.
- [43] Dorota Filipczuk, Charles Mason, and Stephen Snow. Using a Game to Explore Notions of Responsibility for Cyber Security in Organisations. In *Proceedings of CHI 2019 Human Factors in Computing Systems*, pages 1–6, 2019.
- [44] Anders M Fjell, Kristine B Walhovd, Ivar Reinvang, Arvid Lundervold, Anders M Dale, Brian T Quinn, Nikos Makris, and Bruce Fischl. Age does not increase rate of forgetting over weeks—neuroanatomical volumes and visual memory across the adult life-span. *Journal of the International Neuropsychological Society*, 11(1):2–15, 2005.

- [45] Waldo Rocha Flores, Hannes Holm, Marcus Nohlberg, and Mathias Ekstedt. Investigating personal determinants of phishing and the effect of national culture. *Information and Computer Security*, Volume 23:178–199, 2015.
- [46] Vaibhav Garg, Jean Camp, Lesa Mae, and Katherine Connelly. Designing risk communication for older adults. In *Symposium on Usable Privacy and Security (SOUPS)*. Citeseer, 2011.
- [47] Robin Gonzalez and Michael E Locasto. An interdisciplinary study of phishing and spear-phishing attacks. *SOUPS*, 2015.
- [48] Tzipora Halevi, Jim Lewis, and Nasir Memon. Phishing, Personality Traits and Facebook. *arXiv*, 2013.
- [49] Tzipora Halevi, Nasir Memon, and Oded Nov. Spear-Phishing in the Wild: A Real-World Study of Personality, Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks. *SSRN Electronic Journal*, 2015.
- [50] Kyung Wha Hong, Christopher M. Kelley, Rucha Tembe, Emerson Murphy-Hill, and Christopher B. Mayhorn. Keeping Up With The Joneses: Assessing phishing susceptibility in an email task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57:1012–1016, 2013.
- [51] Cristian Iuga, Jason RC Nurse, and Arnau Erola. Baiting the hook: factors impacting susceptibility to phishing attacks. *Human-Centric Computing and Information Sciences*, 6:8, 2016.
- [52] Ankit Jain and BB Gupta. Phishing Detection: Analysis of Visual Similarity Based Approaches. *Security and Communication Networks*, 2017:1–20, 2017.
- [53] Markus Jakobsson, Alex Tsow, Ankur Shah, Eli Blevis, and Youn-Kyung Lim. What instills trust? A qualitative study of phishing. In *Financial Cryptography*, pages 356–361. LNCS, 2007.
- [54] Timothy Kelley and Bennett I Bertenthal. Real-World Decision Making: Logging Into Secure vs. Insecure Websites. *USEC*, 2016.
- [55] Sungjin Kim, Jinkook Kim, and Brent ByungHoon Kang. Malicious url protection based on attackers' habitual behavioral analysis. *Computers & Security*, 77:790–806, 2018.
- [56] Panagiotis Kintis, Najmeh Miramirkhani, Charles Lever, Yizheng Chen, Rosa Romero-Gómez, Nikolaos Pitropakis, Nick Nikiforakis, and Manos Antonakakis. Hiding in plain sight: A longitudinal study of com-bosquatting abuse. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 569–586, 2017.
- [57] Iakovos Kirlappos and Martina A Sasse. Security education against phishing: A modest proposal for a major rethink. *IEEE Security & Privacy*, 10(2):24–32, 2012.
- [58] Iakovos Kirlappos and Martina Angela Sasse. Security Education against Phishing: A Modest Proposal for a Major Rethink. *Security and Privacy*, 10(2):24–32, 2012.
- [59] Sabina Kleitman, Marvin KH Law, and Judy Kay. It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. *PLOS ONE*, 13:e0205089, 2018.
- [60] Ponnurangam Kumaraguru, Justin Cranshaw, and Alessandro Acquisti. School of phish: a real-world evaluation of anti-phishing training. *SOUPS*, 2009.
- [61] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System. In *CHI*, pages 905–914. ACM, 2007.
- [62] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)*, 10:7, 2010.
- [63] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching Johnny Not to Fall for Phish. *Transactions on Internet Technology*, 10(2):1–7, ACM, 2010.
- [64] Alexandra Kunz, Melanie Volkamer, Simon Stockhardt, Sven Palberg, Tessa Lottermann, and Eric Piegert. Nophish: Evaluation of a web application that teaches people being aware of phishing attacks. In *Informatik*, pages 15–24. GI, GI, LNI, 2016.
- [65] Elmer Lastdrager, Ines C Gallardo, Pieter Hartel, and Marianne Junger. How effective is anti-phishing training for children? *Symposium on Usable Privacy and Security (SOUPS)*, 2017.
- [66] Elaine Lau and Zachary NJ Peterson. A research framework and initial study of browser security for the visually impaired. In *Symposium on Usable Privacy and Security (SOUPS)*, 2015.
- [67] D. Lewalter. Cognitive strategies for learning from static and dynamic visuals. *Learning and Instruction*, 13(2):177 – 189, 2003.

- [68] Peter Likarish, Donald E Dunbar, Juan Pablo Hourcade, and Eunjin Jung. Bayeshield: conversational anti-phishing user interface. In *SOUPS*, volume 9, pages 1–1, 2009.
- [69] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does domain highlighting help people identify phishing sites? In *Proceedings of CHI 2011 Human Factors in Computing Systems*, pages 2075–2084. ACM, 2011.
- [70] Gang Liu, Guang Xiang, Bryan A Pendleton, Jason I Hong, and Wenyin Liu. Smartening the crowds: computational techniques for improving human verification to fight phishing scams. *Symposium on Usable Privacy and Security (SOUPS)*, 2011.
- [71] Hans-Henning Lühr. Warm welcome to the website of the it planning council, 2019. https://www.it-planungsrat.de/EN/home/home_node.html;jsessionid=C4D654065D8D53558E2CDDD6287D957B.1_cid332 accessed: 11.09.2019.
- [72] Samuel Marchal, Giovanni Armano, Tommi Grondahl, Kalle Saari, Nidhi Singh, and N. Asokan. Off-the-Hook: An Efficient and Usable Client-Side Phishing Prevention Application. *IEEE Transactions on Computers*, 66, 2017.
- [73] Claudio Marforio, Ramya Jayaram Masti, Claudio Soriente, Kari Kostiainen, and Srdjan Čapkun. Evaluation of personalized security indicators as an anti-phishing mechanism for smartphone applications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 540–551. ACM, 2016.
- [74] Charlie Marriott. *Through the Net Investigating How User Characteristics Influence Susceptibility to Phishing*. PhD thesis, Dublin Institute of Technology, 2018.
- [75] Jaclyn Martin. *Something Looks Phishy Here: Applications of Signal Detection Theory to Cyber-Security Behaviors in the Workplace*. PhD thesis, University of South Florida, 2017.
- [76] Jaclyn Martin, Chad Dubé, and Michael D Covert. Signal Detection Theory (SDT) Is Effective for Modeling User Behavior Toward Phishing and Spear-Phishing Attacks. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 60:1179–1191, 2018.
- [77] Max-Emanuel Maurer and Dennis Herzner. Using visual website similarity for phishing detection and reporting. *Conference on Human Factors in Computing Systems (CHI)*, 2012.
- [78] Peter Mayer, Christian Schwartz, and Melanie Volkamer. On the systematic development and evaluation of password security awareness-raising materials. In *Proceedings of the 34th Annual Computer Security Applications Conference*, pages 733–748. ACM, 2018.
- [79] Christopher B Mayhorn and Patrick G Nyeste. Training users to counteract phishing. *Work (Reading, Mass.)*, 41 Suppl 1:3549–52, 2012.
- [80] Jamshaid G Mohebzada, Ahmed El Zarka, Arsalan H Bhojani, and Ali Darwish. Phishing in a university community: Two large scale phishing experiments. In *2012 International Conference on Innovations in Information Technology (IIT)*, pages 249–254. IEEE, 2012.
- [81] Gregory D Moody, Dennis F Galletta, and Brian Dunn. Which phish get caught? An exploratory study of individuals’ susceptibility to phishing. *European Journal of Information Systems*, 26:564–584, 2017.
- [82] María M Moreno-Fernández, Fernando Blanco, Pablo Garaizar, and Helena Matute. Fishing for phishers. Improving Internet users’ sensitivity to visual deception cues to prevent electronic fraud. *Computers in Human Behavior*, 69:421–436, 2017.
- [83] Steven J Murdoch and Martina A Sasse. Should you phish your own employees? <https://www.benthamsgaze.org/2017/08/22/should-you-phish-your-own-employees/>, 2017. accessed: 18.09.2019.
- [84] Stephan Neumann, Benjamin Reinheimer, and Melanie Volkamer. Don’t be deceived: the message might be fake. In *International Conference on Trust and Privacy in Digital Business*, pages 199–214. Springer, 2017.
- [85] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 6412–6424, 2017.
- [86] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. Phishing for the Truth: A Scenario-Based Experiment of Users’ Behavioural Response to Emails. pages 366–378, 2013.
- [87] M Pattinson, C Jerram, K Parsons, A McCormac, and M Butavicius. Managing Phishing Emails: A Scenario-Based Experiment. pages 74–85, 2011.

- [88] Justin Petelka, Yixin Zou, and Florian Schaub. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings . In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[89] Richard Roberts, Yaelle Goldschlag, Rachel Walter, Taejoong Chung, Alan Mislove, and Dave Levin. You are who you appear to be: A longitudinal study of domain impersonation in tls certificates. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2489–2504, 2019.

[90] Dawn M Sarno, Joanna E Lewis, Corey J Bohil, Mindy K Shoss, and Mark B Neider. Who are Phishers luring?: A Demographic Analysis of Those Susceptible to Fake Emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61:1735–1739, 2017.

[91] Stuart E Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor’s new security indicators. In *2007 IEEE Symposium on Security and Privacy (SP’07)*, pages 51–65. IEEE, 2007.

[92] Steve Sheng, Mandy Holbrook, Ponnurangam Ku-maraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of CHI 2010 Human Factors in Computing Systems*, pages 373–382. ACM, 2010.

[93] Steve Sheng, Bryant Magnien, Ponnurangam Ku-maraguru, Alessandro Acquisti, Lorrie Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. *Symposium on Usable Privacy and Security (SOUPS)*, 2007.

[94] Steve Sheng, Bryant Magnien, Ponnurangam Ku-maraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 88–99. ACM, 2007.

[95] Akbar Siami Namin, Rattikorn Hewett, Keith S Jones, and Rona Pogrund. Sonifying internet security threats. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2306–2313. ACM, 2016.

[96] Gunikhan Sonowal, KS Kuppusamy, and Ajit Kumar. Usability evaluation of active anti-phishing browser extensions for persons with visual impairments. *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 1–6, 2017.

[97] Harold Stanislaw and Natasha Todorov. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1):137–149, 1999.

[98] Simon Stockhardt, Benjamin Reinheimer, Melanie Volkamer, Peter Mayer, Alexandra Kunz, Philipp Rack, and Daniel Lehmann. Teaching Phishing-Security: Which Way is Best? *IFIP SEC*, pages 135–149, 2016.

[99] Janos Szurdi, Balazs Kocso, Gabor Cseh, Jonathan Spring, Mark Felegyhazi, and Chris Kanich. The long “tail” of typosquatting domain names. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 191–206, 2014.

[100] Christopher Thompson, Martin Shelton, Emily Stark, Maximilian Walker, Emily Schechter, and Adrienne Porter Felt. The web’s identity crisis: understanding the effectiveness of website identity indicators. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1715–1732, 2019.

[101] Tom N Tombaugh and Anita M Hubley. Rates of forgetting on three measures of verbal learning using retention intervals ranging from 20 min to 62 days. *Journal of the International Neuropsychological Society*, 7(1):79–91, 2001.

[102] Kai Florian Tschakert and Sudsanguan Ngamsuri-yaroj. Effectiveness of and user preferences for security awareness training methodologies. *Heliyon*, 5:e02010, 2019.

[103] Anthony Vance, Brock Kirwan, Daniel Bjorn, Jeffrey Jenkins, and Bonnie Anderson. What Do We Really Know about How Habituation to Warnings Occurs Over Time?: A Longitudinal fMRI Study of Habituation and Polymorphic Warnings. In *Proceedings of CHI 2017 Human Factors in Computing Systems*, pages 2215–2227, 2017.

[104] Melanie Volkamer, Karen Renaud, and Paul Gerber. Spot the phish by checking the pruned URL. *Information and Computer Security*, Volume 24:372–385, 2016.

[105] Melanie Volkamer, Karen Renaud, and Benjamin Reinheimer. Torpedo: Tooltip-powered phishing email detection. In *IFIP SEC*, pages 161–175. Springer, 2016.

[106] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of torpedo: Tooltip-powered phishing email detection. *Computers & Security*, 2017.

- [107] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, Philipp Rack, Marco Ghiglieri, Peter Mayer, Alexandra Kunz, and Nina Gerber. Developing and Evaluating a Five Minute Phishing Awareness Video. *Trust, Privacy and Security in Digital Business (Trust-Bus)*, pages 119–134, 2018.
- [108] Melanie Volkamer, Simon Stockhardt, Steffen Bartsch, and Michaela Kauer. Adopting the cmu/apwg anti-phishing landing page idea for germany. In *STAST*, pages 46–52. IEEE, 2013.
- [109] Jingguo Wang, Yuan Li, Columbia College, H Raghav Rao, and The University of Texas at San Antonio. Overconfidence in Phishing Email Detection. *Journal of the Association for Information Systems*, 17:759–783, 2016.
- [110] Rick Wash and Molly Cooper. Who Provides Phishing Training? In *Proceedings of CHI 2018 Human Factors in Computing Systems*, 2018.
- [111] Zikai Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. page 108, 2019.
- [112] Zikai Alex Wen, Yiming Li, Reid Wade, Jeffrey Huang, and Amy Wang. What.Hack: Learn Phishing Email Defence the Fun Way. pages 234–237, 2017.
- [113] Min Wu, Robert C Miller, and Simson L Garfinkel. Do security toolbars actually prevent phishing attacks? *CHI*, pages 601–610, 2006.
- [114] Weining Yang, Jing Chen, Aiping Xiong, Robert W Proctor, and Ninghui Li. Effectiveness of a phishing warning in field settings. *the 2015 Symposium and Bootcamp*, page 14, 2015.
- [115] Ka-Ping Yee. Designing and evaluating a petname anti-phishing tool. In *Poster presented at Symposium on usable Privacy and Security (SOUPS)*, pages 6–8, 2005.
- [116] Tianjian Zhang. Knowledge Expiration in Security Awareness Training. *Conference on Digital Forensics, Security and Law (ADFSL)*, 2018.
- [117] Yue Zhang, Serge Egelman, Lorrie Faith Cranor, and Jason Hong. Phinding Phish: Evaluating Anti-Phishing Tools. In *NDSS*. School of Computer Science, Internet Society, 2007.
- [118] Yue Zhang, Jason I Hong, and Lorrie F Cranor. CANTINA: A Content-Based Approach to Detecting Phishing Web Sites. In *16th International World Wide Web Conference*, pages 639–648, 2007.
- [119] Olga A Zielinska, Rucha Tembe, Kyung Hong, Xi Ge, Emerson Murphy-Hill, and Christopher B Mayhorn. One Phish, Two Phish, How to Avoid the Internet Phish. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58:1466–1470, 2014.

A Reminder Measures

How to Detect Fraudulent and Phishing Mails

General Information

Criminals use various strategies to harm you. Popular attack strategies are:

- the dissemination of malware to e.g. gain access to your devices or
- deception of the end users to obtain sensitive information (e.g. access data).

A widely used attack method is to send fraudulent messages to you that pretend to have a legitimate reason. Fraudulent messages may be received via different channels, e.g. as emails, SMSs, via Messenger or social networks.

The contents of these messages may be dangerous in different ways:

Sensitive data: the messages ask you to return sensitive data, such as access data or documents worth protecting.

Money transfers/calls: messages ask you to transfer money or to call e.g. cooperation partners, supposed friends or business partners. In this way, criminals will get the money from you by direct transfer or the money is debited with the telephone invoice.

Links: messages may contain one or several dangerous links. The fraud is aimed at making you click one of these links. These links will then lead you to e.g. a deceptively real-looking, but fraudulent website (also called phishing site) where you are supposed to log in.

Alternatively, you are guided to a website that installs malware on your device.

Attachments: messages contain one or several dangerous files (e.g. an attachment of an email). The criminals want to make you open the attachment. By opening or executing the file, malware is installed on your device.

Advertisements: messages may contain ads or other worthless contents (these messages are frequently called spams). The attack is aimed at making you buy something. In reality, the primary damage is lost working time, because you look at the message, assess it, and delete it.

Figure 6: Excerpt of the *text* reminder measure.



Figure 7: An impression of the video reminder measure.

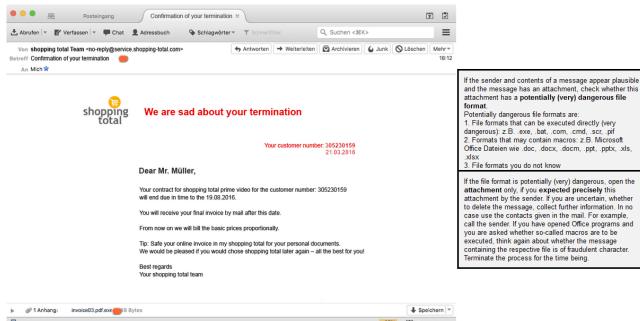


Figure 8: One of the interactive example measures used. The red dots represent the interactivity-points where participants can reveal more information about the respective area of the email.

Fraudulent Mails

1. Rule: Check the sender and the contents of every mail for plausibility.

- ✗ The sender shop@[sye.jp](#) for an Amazon email
- ✓ The sender rechnung@[amazon.com](#) for an Amazon email

2. Rule: Get familiar with where to find the real web address behind a link (e.g. for PCs or laptops in the tooltip or the status bar).

3. Rule: Identify the whois of the web address.

<https://nophish.amazon.com/login>

4. Rule: Check, if the whois matches with the supposed legit mail.

- ✓ <https://www.my-parcelservice.com/>
- ✗ <https://www.my-parcelservice.com.online-shopping.com/>
- ✗ <https://online-shopping.com/my-parcelservice.com/>
- ✗ <https://129.13.152.9/secuso.org.secuso.org.secure-login.com/>

5. Rule: Check, if the whois is written correctly.

- ✓ <https://www.farmers-market-total.com/>
- ✗ <https://www.farmers-market-total.com/>
- ✗ <https://www.farmers-market-total.com/>
- ✗ <https://www.farmers-market-total.com/>

6. Rule: If you cannot assess the whois clearly, collect further information, e.g. by searching the address in a search machine.

- ✓ <https://www.amazon.com/>
- ✗ <https://www.amazon-shopping.com/>

7. Rule: Check the file format of the attachment.

- ✗ Executable formats, e.g. .exe, .bat, .cmd
- ✗ Files including macros, e.g. Office files like .doc, .docx, .doxm

8. Rule: If you cannot clearly assess the attachment or if you are uncertain about expecting precisely this format by the sender, collect further information, e.g. by contacting the sender. In no case use the contacts given in the mail.

Figure 9: The shortened text measure.

B Attack Types

(a) Phishers may try to trick recipients by displaying the legitimate URL in the email's message text (and hope that recipients do not check the actual link destination). (b) Phishers may try to trick recipients by ‘scripting’ a tooltip with the legitimate URL next to the link. (c) Phishers replace the legitimate URL with a domain that they own (which has no connection to the expected domain). They might adopt either the sub-domain (e.g., <https://www.amazon.com.phisher.com>) and/or the path so that the expected domain appears to allay suspicions. (d) Phishers replace the legitimate URL by a link to a domain that they own, and which looks very similar to the expected one (e.g., arnazon.de). (e) Phishers replace the legitimate URL by one with a domain they own, and which extends the expected domain name — most likely a word before or after the original name is added (e.g. amazon-secure.com).

C Study Scenario

D Result Related Information

| Sensitivity (d') | | | | |
|------------------|----------|----------|----------|----------|
| M_{pre} | M_{0M} | M_{4M} | M_{6M} | M_{8M} |
| 1.11 | 2.13 | 1.60 | 1.45 | 1.39 |
| Criterion (C) | | | | |
| 0.12 | -0.05 | .09 | -0.02 | 0.15 |

Table 1: SDT mean results for RQ1

- You are Martin Müller and have two e-mail addresses: martin.mueller.77@web.de and martin.mueller.77@posteo.de
- You speak German and English
- Your boss is Wolfgang Lange. His e-mail address is wolfgang.lange.54@web.de
- Your colleague is Jonas Schmidt. He has two e-mail addresses: jonas.schmidt.77@web.de and jonas.schmidt.77@posteo.de
- You use all the services used in this evaluation
- You use the various operating systems and programs used in this evaluation

Figure 10: Study details for role play (translated from German for the purpose of this paper).

– Continued on next page –

D.1 Results for Research Question 1

For the reporting of the results, we combine the analyses of $H_{M_{pre}-M_{0M}}$ and $H_{M_{pre}-M_{\Delta t M}}$ (for $\Delta t = 4 + 2i$ months, where $i \in \{0, 1, 2\}$)

We checked for a significant difference between the corresponding five groups (see Figure 1) using a one-way ANOVA. There was statistical significance between the groups ($F(4, 227) = 5.457, p < 0.001$) for the sensitivity (d'). For the effect size we calculated $\omega^2 = .093$, which is a medium effect size according to [42]. A LSD post-hoc showed that the sensitivity for the M_{0M} ($d' = 2.13, SD = 1.15$) was significantly higher than for the M_{pre} ($d' = 1.11, SD = 1.12$) with ($p < .001$).

The LSD post-hoc test showed that the sensitivity for the M_{4M} ($d' = 1.60, SD = 1.01$) was significantly higher than for the M_{pre} ($d' = 1.11, SD = 1.12$) with ($p = .034$). But the sensitivity for the M_{6M} ($d' = 1.46, SD = 1.01$) with ($p = .123$) and for the M_{8M} with ($d' = 1.39, SD = 1.42$) with ($p = .155$) was not significantly higher.

Note, there was no statistical significance between the groups ($p = 0.623$) for the criterion (C).

In summary: We accept $H_{M_{pre}-0Month}$ and $H_{M_{pre}-4Months}$.

D.2 Results for Research Question 2

First, we checked for which reminder measures the hypothesis $H_{M_{pre}-M_{Reminder_x-6M}}$ holds.

We checked for a significant difference between M_{pre} and the four months retention groups (see Figure 1) using a one-way ANOVA. There was statistical significance between the groups ($F(5, 244) = 2.410, p = 0.037$) for the sensitivity (d'). For the effect size we calculated $\omega^2 = .027$, which is a small effect size [42]. A LSD post-hoc showed that the sensitivity for the $M_{Reminder-Text-6M}$ ($d' = 1.61, SD = 1.18$) with ($p = .005$), $M_{Reminder-Video-6M}$ ($d' = 1.80, SD = 1.42$) with ($p = .005$) and $M_{Reminder-InteractiveExamples-6M}$ ($d' = 1.73, SD = 1.19$) with ($p = .007$) were significantly higher than for the M_{pre} ($d' = 1.11, SD = 1.12$). The sensitivity for the $M_{Reminder-ShortText-6M}$ was not significantly higher ($d' =$

1.56, $SD = 1.11$) with ($p = .075$). Note, there was statistical significance between the groups ($p = 0.013$) for the criterion (C). A LSD post-hoc showed that the criterion for the $M_{Reminder-Text-6M}$ ($C = -.23, SD = .59$) with ($p = .043$) and $M_{Reminder-InteractiveExamples-6M}$ ($C = -.43, SD = .65$) with ($p < .001$) were significantly different from the M_{pre} ($C = .12, SD = .84$). The criterion for the $M_{Reminder-ShortText-6M}$ ($C = .03, SD = .77$) with ($p = .603$) and $M_{Reminder-Video-6M}$ ($C = -.06, SD = .70$) with ($p = .273$) was not significantly different.

In summary: We accept $H_{M_{pre}-M_{Reminder_x-6M}}$ for text measure, video measure, and interactive examples measure.

In order to test whether one of the three remaining reminder measures performs best, we also checked the ANOVA values for between the reminder measures. There is no significant difference between these measures. From the descriptive data, the interactive examples measure performs slightly better than the video measure (see Figure 5).

D.3 Results for Research Question 3

Based on the results from RQ2 we decided to not collect data from the short text group after 12 months. In order to address the pre-conditions from Section 5.1 we kind of extended $H_{M_{pre}-M_{Remindery-12M}}$ accordingly, i.e. six measurements were considered (see Figure 2).

We linked participants using the provided codes. This resulted in 20 participants in M_{6M-12M} , 17 participants in $M_{Reminder-Text-12M}$, 17 participants in $M_{Reminder-Video-12M}$, and 12 participants in $M_{Reminder-InteractiveExamples-12M}$.

We analysed the data from participants that we could link via code. We checked for a significant difference between the corresponding six measurements. There was statistical significance between the groups ($F(5, 172) = 2.721, p = 0.022$) for the sensitivity (d'). The LSD post-hoc showed that the sensitivity for the $M_{Reminder-Text-12M}$ ($d' = 1.93, SD = 1.17$) with ($p = .009$), the $M_{Reminder-Video-12M}$ ($d' = 1.77, SD = 1.32$) with ($p = .031$) and $M_{Reminder-InteractiveExamples-12M}$ ($d' = 1.96, SD = 1.34$) with ($p = .016$) were significantly higher than for the M_{pre} ($d' = 1.11, SD = 1.12$). The sensitivity for the M_{12M} ($d' = 1.55, SD = 1.06$) with ($p = .060$) was not significantly higher. The sensitivity for the M_{6M-12M} ($d' = 1.43, SD = 0.77$) with ($p = .256$) was not significantly higher. For the effect size we calculated $\omega^2 = .047$, which is a small effect size according to [42]. Note, there was no statistical significance between the measurements ($p = 0.274$) for the criterion (C).

| Sensitivity (d') | | | | | |
|------------------|------------------------|-------------------------|---------------------------------------|-----------------------------|--|
| M_{pre} | $M_{Reminder-Text-6M}$ | $M_{Reminder-Video-6M}$ | $M_{Reminder-InteractiveExamples-6M}$ | $M_{Reminder-ShortText-6M}$ | |
| 1.11 | 1.61 | 1.73 | 1.80 | 1.56 | |
| Criterion (C) | | | | | |
| 0.12 | -0.02 | -0.23 | -0.43 | -0.06 | |

Table 2: SDT mean results for RQ2

| Sensitivity (d') | | | | | | |
|------------------|-----------|-----------|--------------|-------------------------|--------------------------|--|
| Measure | M_{pre} | M_{12M} | M_{6M-12M} | $M_{Reminder-Text-12M}$ | $M_{Reminder-Video-12M}$ | $M_{Reminder-InteractiveExamples-12M}$ |
| Linked | 1.11 | 1.55 | 1.44 | 1.86 | 1.77 | 1.96 |
| Unmatched | 1.11 | 1.55 | 1.60 | 1.54 | 1.73 | 1.69 |
| Criterion (C) | | | | | | |
| Linked | 0.12 | 0.12 | 0.12 | -0.07 | -0.05 | -0.38 |
| Unmatched | 0.12 | 0.12 | 0.01 | -0.08 | 0.00 | -0.02 |

Table 3: SDT mean results for RQ3

| Measure | M_{12M} | M_{6M-12M} | $M_{Reminder-Text-12M}$ | $M_{Reminder-Video-12M}$ | $M_{Reminder-InteractiveExamples-12M}$ |
|---------|-----------|--------------|-------------------------|--------------------------|--|
| Linked | 36 | 20 | 17 | 17 | 12 |

Table 4: Number of participants per measure for linked and Unmatched groups

| Minutes | $M_{Reminder-Text-6M}$ | $M_{Reminder-Video-6M}$ | $M_{Reminder-InteractiveExamples-6M}$ | $M_{Reminder-ShortText-6M}$ |
|----------------|------------------------|-------------------------|---------------------------------------|-----------------------------|
| Median | 9.93 | 9.50 | 7.32 | 3.35 |
| Std. Deviation | 7.52 | 3.68 | 12.81 | 15.75 |

Table 5: The time needed by participants

E Email Screenshots of Phishes

| Reason | Phishing Email Screenshot |
|----------------------------------|---------------------------|
| Implausible Content and Document | |
| Random URL | |

Figure 11: Phishing email screenshots (part 1)

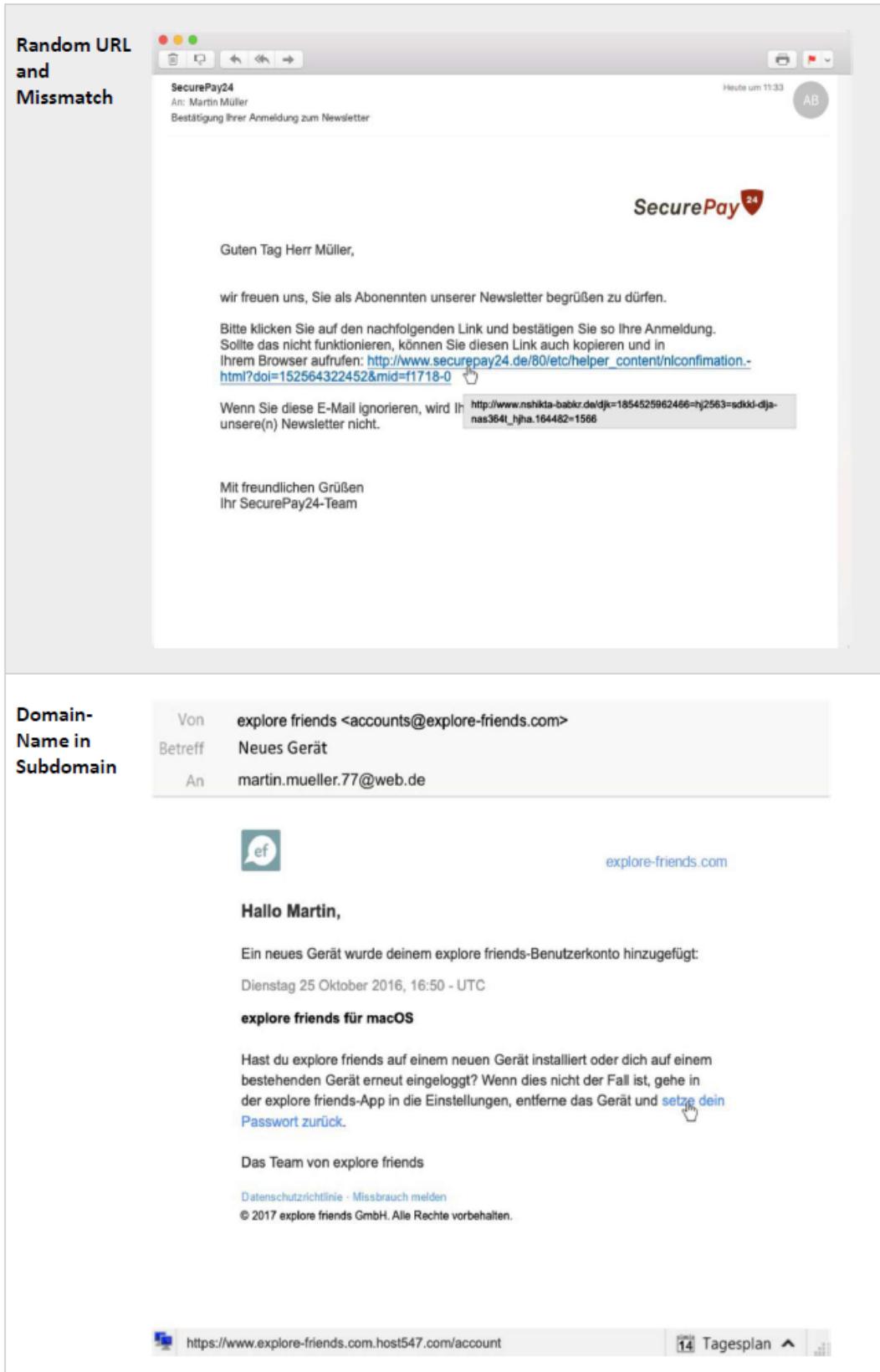


Figure 12: Phishing email screenshots (part 2)

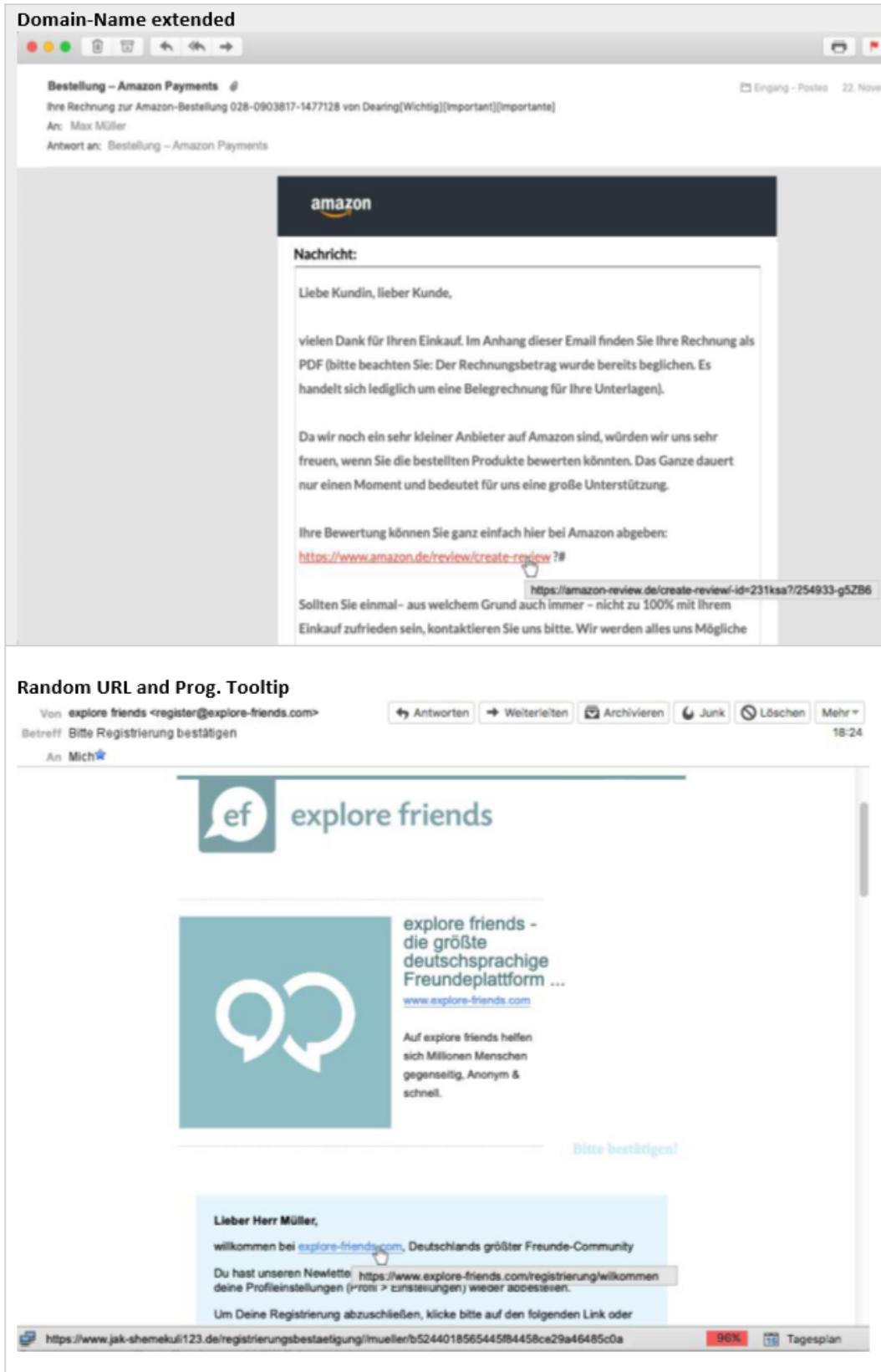


Figure 13: Phishing email screenshots (part 3)

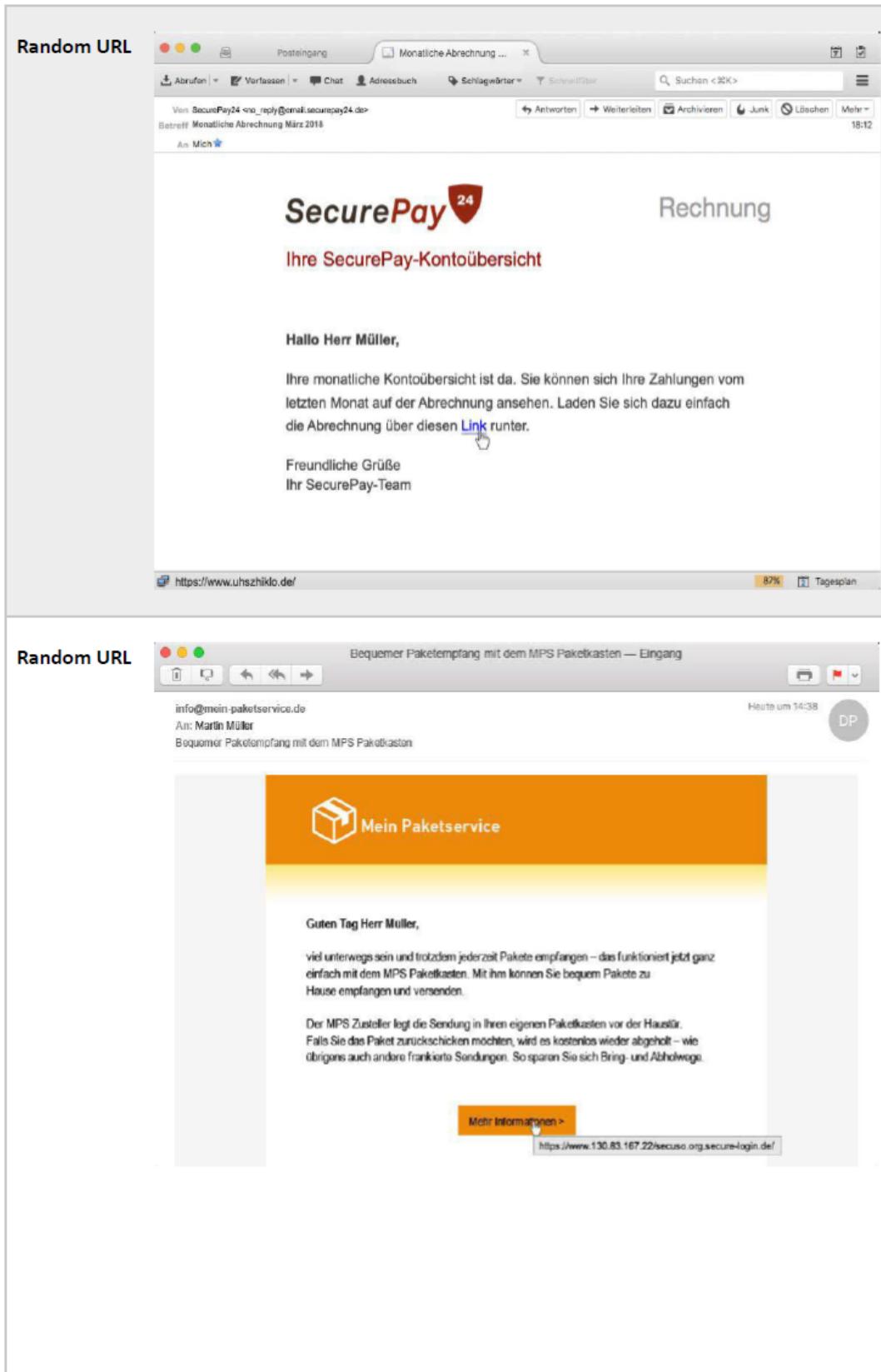


Figure 14: Phishing email screenshots (part 4)

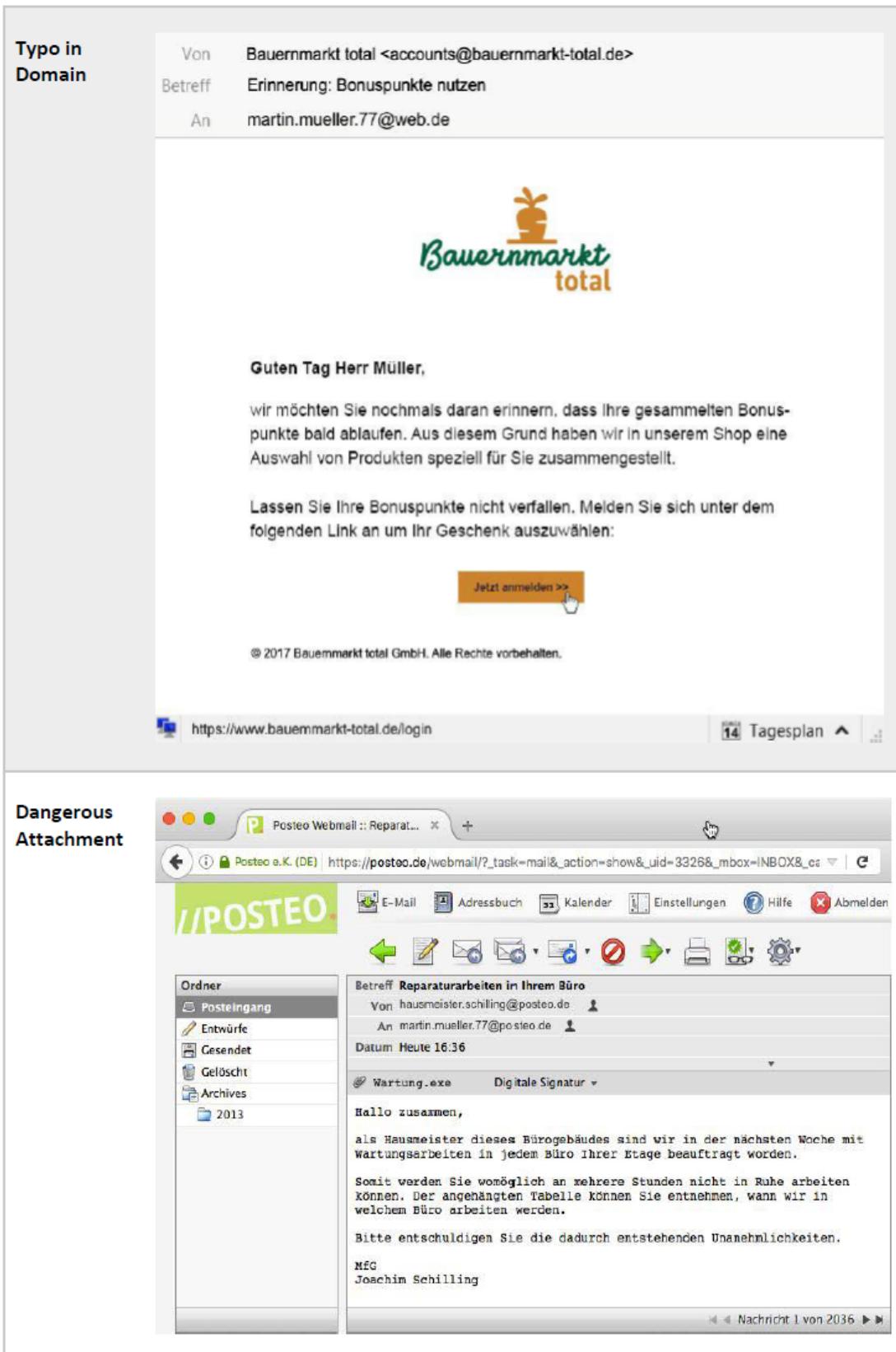


Figure 15: Phishing email screenshots (part 5)