Better the Phish You Know: Evaluating Personalization in Anti-Phishing Learning Games

Rene Roepke[1,*], Vincent Drury[2,*], Ulrike Meyer[2] and Ulrik Schroeder[1]
[1]*Learning Technologies Research Group, RWTH Aachen University, Ahornsstr. 55, 52074 Aachen, Germany*
[2]*IT-Security Research Group, RWTH Aachen University, Mies-v.-d.-Rohe-Str. 15, 52074 Aachen, Germany*

Keywords: Anti-Phishing Education, Game-based Learning, Personalization, User Study, Gameplay Analysis.

Abstract: Anti-phishing learning games present a motivating, interactive approach to user education and thus, various games have been developed and studied in the past. A common trend among these games is a limited use of game mechanics and no consideration of learners using methods of personalization. In this paper, we compare an anti-phishing learning game with its personalized version in the scope of a longitudinal user study with 89 participants. For personalization, the player's familiarity with different services is used to provide personalized content in the form of URLs in the game. To further understand the effects of personalization, we analyze game log data and evaluate how players interact with personalized learning game content. While the comparison of both game versions did not yield significant differences in the participants' performance in URL tests, the in-game analysis confirmed that players interact differently when confronted with URLs based on services they are not familiar with compared to those they use or know. These differences when handling unknown URLs in the in-game analysis might indicate, that personalization could be leveraged to improve awareness and the knowledge transfer to the real world.

1 INTRODUCTION

A common threat to Internet users worldwide is phishing, "a scalable act of deception whereby impersonation is used to obtain information from a target" (Lastdrager, 2014). Current trend reports observe high numbers of newly created phishing websites (APWG, 2021) as well as clicks on phishing links (Kaspersky, 2021). While phishers employ a diverse repertoire of attack vectors, including email, instant messaging, and even voice phishing (Aleroud and Zhou, 2017), these trend reports indicate that links to phishing websites still present an imminent threat to users. Teaching users to recognize potentially malicious URLs, and therefore phishing websites and malicious links, can help alleviate the problem. Therefore, researchers have explored different approaches to user education, ranging from traditional awareness campaigns to user training using simulated phishing attacks or game-based learning.

While various anti-phishing learning games have been proposed in the past, a common trend of existing games seems to be the use of limited game mechanics and failing to consider the learner by means of personalization (Roepke et al., 2020a). With phishing being an imminent threat, users will be presented with various phishing messages claiming to be from services they know and those they do not know. Depending on which case, users can apply different strategies to recognize phishing and protect themselves. As existing anti-phishing do not yet consider the learners' familiarity with different services, they fail to reflect this situation, which presents a research gap in the field of game-based anti-phishing education.

Considering the learners' familiarity with services in anti-phishing learning games can enable new approaches for elaborated feedback or adaptive gameplay to support the learning experience. Furthermore, the use of more relevant services or a more realistic decision strategy, which considers the learners' familiarity with a service, might have a positive impact on their awareness in a real-world attack. To achieve this, personalization needs to be implemented, e.g., using the conceptual approach and framework for personalization of anti-phishing learning games as presented in (Roepke et al., 2021b). Consequently, these implementations need to be compared with traditional, non-personalized games to better understand the advan-

* These authors contributed equally.

Roepke, R., Drury, V., Meyer, U. and Schroeder, U.
Better the Phish You Know: Evaluating Personalization in Anti-Phishing Learning Games.
DOI: 10.5220/0011042100003182
In *Proceedings of the 14th International Conference on Computer Supported Education (CSEDU 2022) - Volume 2*, pages 458-466
ISBN: 978-989-758-562-3; ISSN: 2184-5026
Copyright © 2022 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

tages and benefits of personalization. An exploratory analysis of gameplay using detailed event log data would allow even more insights into personalization and its effects on players.

In this paper, we present a comparative user study (N=89) of an existing learning game and its personalized version in a pre-/post-test design. The used game prototype was previously presented in (Roepke et al., 2021a) and personalized using the personalization framework presented in (Roepke et al., 2021b). Additional longitudinal tests ($N = 36$) as well as an in-game analyses of the participants' gameplay ($N = 49$) allow further insights into how personalization affects the participants' performance and behavior. While the results of our comparison of the game's personalized and non-personalized version in the post-test are inconclusive, in that personalization did not outperform the traditional version of the game, the analysis of in-game behavior using game log data revealed differences in players' actions. As expected, the results show that players classification accuracy differs for different levels of familiarity, i.e. players show difficulties when classifying URLs of unknown services. We therefore demonstrate, that there are definite advantages to using the personalized version, and propose several possible venues for future research.

2 RELATED WORK

This paper describes a comparative study of a personalized anti-phishing learning game with a non-personalized version and explores the effects of game content personalization. In prior studies on game-based anti-phishing education, different approaches have been evaluated and the effectiveness of games for anti-phishing education has been shown for different user groups (Sheng et al., 2007; Canova et al., 2015; Drury et al., 2022). However, existing games have been criticized as their design may limit the potential learning outcomes and does not consider the learners and their familiarity with the learning content, i.e. personalizing presented URLs which have to be classified as either malicious or benign within the game (Roepke et al., 2020a). So far, personalization of anti-phishing learning games has not been explored or even implemented for possible evaluation in user studies. There are, however, other types of anti-phishing educational material that have explored personalization or customization. In particular, researchers have taken a look at spear phishing, a more sophisticated type of phishing that is tailored towards a recipient, and whether customized training can help prevent it. In (Kumaraguru et al., 2008), the embedded training against spear phishing was explored, showing that customized content led to an advantage when detecting spear phishing attacks compared to regular educational material.

Beyond anti-phishing learning games, personalization of games has been the subject of different research projects (Law and Rust-Kickmeier, 2008; Kickmeier-Rust and Albert, 2010). Here, adaptivity on a micro or macro level has been implemented to provide personalized storytelling or dynamic difficulty adjustment, i.e. adaptive gameplay which matches difficulty to players' skill level. Personalization through adaptivity focuses more on sequencing and structuring of learning content and less on actual adaptation of the content itself. As we did not find any projects using game content personalization, the respective research area still has a potential to be explored. However, outside the educational domain, research on game content generation (Dey and Konert, 2016) may provide interesting approaches.

Since neither implementation nor evaluation of personalized anti-phishing learning games has been done prior to this work and existing games fail to consider individual learners (Roepke et al., 2020b), we identify an untapped potential for the personalization of anti-phishing learning games to provide a more suitable game-based learning environment which supports learners in their different learning contexts. Furthermore, a comparison to existing non-personalized games could yield meaningful insights regarding the effectiveness of personalized games. With recent work introducing a concept (Roepke et al., 2020b) as well as an implementation of personalization framework for anti-phishing learning games (Roepke et al., 2021b), the natural next step is to conduct a study comparing personalized and non-personalized versions of a game. In addition, an exploratory analysis of gameplay may provide insights into players' behavior when dealing with different learning content.

3 STUDY SETUP

For the comparison of a personalized and a non-personalized version of a learning game, we chose a between-group design in a pre-/post-test setup including an additional longitudinal test. The study was performed in two batches: the first group played the non-personalized game in November 2020, and another group of participants played the personalized version in May 2021. While the games serve as the independent variables, the performance and confidence in pre-, post- and longitudinal tests serve as dependent variables. This allows for a comparison of the

effect of personalization as well as the exploration of in-game behavior in the personalized game. Additionally, the results of the longitudinal study were analyzed to gain insights into knowledge retention as well as several self-reported characteristics of the participants after playing either one of the games. Our study was therefore designed to answer the following research questions (RQs):

1. How does personalization affect the participants' performance/confidence in classifying URLs?
2. How do the participants' performances change in pre-, post- and longitudinal test?
3. How does personalization (i.e. familiarity with services) affect in-game behavior?

3.1 Games and Personalization

For the main intervention in this study we used the learning game prototype "All sorts of Phish" presented in (Roepke et al., 2021a). In the following, we refer to it as the *analysis game*[1]. The analysis game teaches the basics of the URL structure and different manipulation techniques used to create malicious URLs and deceive users in phishing attacks. The URL structure is explained by introducing three main parts: subdomain, registrable domain, and path. For each part, different manipulation techniques are presented to understand how phishers create malicious and deceiving, but also valid URLs.

The game utilizes a sorting mechanic where players have to analyze and classify given URLs into different categories by sorting them into different buckets (see Figure 1). Each bucket represents a specific URL category derived from applied manipulation techniques for phishing URLs. The categories are based on the URL structure and indicate, where the original domain or deceptive keyword is present. As such, the considered categories are: "IP", "Random", "RegDomain", "Subdomain", and "Path". Furthermore, buckets for benign URLs ("No-Phish") and for discarding unknown URLs ("No idea") are available. The more elaborate sorting mechanic extends the state-of-the-art as most games rely on a binary decision scheme in which players only classify URLs as benign or malicious. The extended classification allows for more insights into the decision process and can reveal players' misconceptions (e.g., by analyzing classification outcomes for different URL categories; see Section 5.2).

In our approach to extend current state of the art, we adapted the analysis game by utilizing a person-

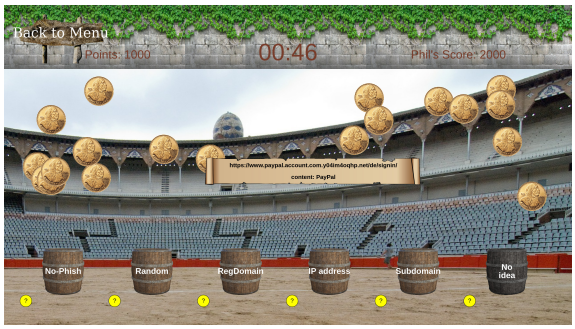

Figure 1: Level of "All sorts of Phish". Players have to classify given URLs, which are hidden behind coins.

alization framework to provide personalized learning game content (Roepke et al., 2021b). The new version of the game is referred to as *personalized game*. The framework first provides a selection interface for players to select services they either use, know but do not use, or do not know at all from a set of services (e.g. "PayPal", "eBay"). The players' selection is then used to compute a learner model, an abstract representation of the learner's characteristics (Bull, 2004), in this case realized as information about the players' familiarity with different services. Next, URLs for all URL categories are created using a URL generator which applies different manipulation techniques to base URLs of a given services (e.g. manipulation of the registrable domain, subdomain or path). The learner model is used as input to the generator such that a set of URLs for different types of service can be created (i.e. services that players use, know but do not use, or do not know). Generated URLs are then embedded into the game to provide a personalized version of the game for individual players. The game purposefully includes a number of services that are less well known, which are included in the game to understand how participants handle such unknown services. The current version of the game presents known and unknown URLs at a 4:1 ratio. Due to randomness implemented in the rules used by the URL generator, returning players will encounter different URLs compared to previous gameplay sessions. Beyond personalization, the game fully supports event logging of all in-game actions, timings and results.

By utilizing the personalization framework, we are able to create a personalized version for each participant of our study and thus, we are able to compare the personalized and non-personalized versions of the analysis game as well as explore in-game behavior of players of the personalized game using event log data.

[1]https://gitlab.com/learntech-rwth/erbse/analysis-game, online, accessed 2022-02-18

3.2 Procedure

The study was conducted as a remote lab study using video conferencing software and a web browser on participants' devices. It was structured into five phases: (1) For the introduction, participants were briefed about the topic of the study and presented with a definition of phishing. (2) Next, participants were presented with the pre-test part of the survey. (3) After finishing the pre-test, the survey software directed participants to either the analysis game or the personalized game. (4) After playing either one of the games, participants returned to the survey for the post-test. (5) When all participants finished the survey, a debriefing informed the participants about the overall goal of the study and answered open questions before closing the session.

Participants were asked to start the survey and proceed at their own pace, as no further instructions were necessary. In case of questions or if technical support was needed, participants were able to immediately contact the instructors and receive help without disrupting other participants in continuing the study. The participants were not told that different games would be tested, nor did they know which group they were assigned to.

For the longitudinal test, all participants were contacted three months after the original study, and invited to take part within a four-weeks time frame. The longitudinal study did not require additional expert support and only contained a two-part survey, as described in the next section.

3.3 Apparatus and Materials

The following questionnaires were used in the different parts of the study:

- **URL Test**: A test consisting of 20 (pre) and 30 (post, longitudinal) URLs to be classified as either benign or phishing URLs. It also includes a question regarding the participants' confidence in their decision for each URL using a 6-point Likert scale. The test was included to measure the overall effect of the interventions, including the comparison of URLs of familiar and unknown services. For the post-test as well as longitudinal testing, ten additional URLs were provided to check for potential learning bias. A list of all URLs used in the longitudinal test can be found in Table 1, while the URLs used in pre- and post-test can be found in (Drury et al., 2022).
- **Recognition of Services** (post/longitudinal)**:** A questionnaire listing all services that were used to create URLs of the URL tests for participants to select for each service whether they (a) use it, (b) do not use it, but know it, or (c) whether it is unknown to them. This test was included to be able to analyze the effect of familiarity with a service on classification performance and confidence in the URL tests.
- **Demographics:** Questionnaire which is used to collect demographic data, including age, gender and educational background.
- **Behavioral Change Questionnaire:** Consists of nine items about participants' behavior towards phishing after participating in the pre-/post-test part of the study (see Table 2). Dividing the items into four categories (with Cronbach's α reliability) provides insights into self-reported application of knowledge (App, $\alpha = .861$), interest in learning more about security using games (Int, $\alpha = .629$), behavioral change (BC, $\alpha = .830$) and the perception of phishing as a threat (PT, $\alpha = .782$). The items use a 6-point Likert scale (1 = "strongly disagree" to 6 = "strongly agree").

The URLs used in the pre-, post- and longitudinal

Table 1: URLs of **URL Test** in longitudinal test; for URLs used also in pre- and post-test, see (Drury et al., 2022).

URL	Category
https://www.facebook.com/login/device-based/re...	Benign
https://www.dropbox.com/login	Benign
https://www.twitch.tv/	Benign
https://www.45m64or.ru/NZYJolaEiBSOSOC...	Random
https://mobile-support.de/en/auth/login?client_id=0...	RegDomain
https://meine.deutssche-bank.de/?client_id=HyB...	RegDomain
https://www.fodus.de/ajax/login/	RegDomain
https://idealo.de%76%73%6C%38%6A%6D%31...	RegDomain
https://www.commerzbank.de-account.support/...	Subdomain
https://login.live.com.id.online/de/login.exe?to=%...	Subdomain

Table 2: Behavioral Change Questionnaire.

Item	
App1	I have been using the things I learned in the game during the past months.
App2	Since playing the learning game, I have been checking the URLs of websites before I click on them.
App3	Since playing the learning game, I have been checking the URLs of websites before I enter personal data (e.g., account credentials).
Int1	Playing the learning game has raised my interest in phishing or other IT security topics.
Int2	I would like to learn more about phishing or other IT security topics by playing learning games.
BC1	Since playing the learning game, I have become more aware of phishing attacks.
BC2	After playing the learning game, I adapted my behavior in dealing with URLs.
PT1	After playing the learning game, I feel like I can protect myself against phishing attacks.
PT2	After playing the learning game, I feel less likely to fall for phishing attacks.

tests were generated by collecting benign login URLs from popular websites in our country of origin (according to Alexa[2] and Tranco[3]). Then, different manipulation techniques were applied to these benign login URLs to create various phishing URLs. We differentiate these manipulation techniques by which part of the URL contains the original target domain or a deceptive keyword: a subdomain, the registrable domain, the path, or none (random URLs). We further differentiate URLs that contain an IP address as host from other URLs with a deceptive part in the path. In all, 13 phishing and 7 benign URLs were created for the pre-test, with 7 phishing and 3 benign URLs added in the post- and longitudinal tests respectively to control for learning bias of the pre-test URLs. While all participants were shown the same URLs as part of the URL test, the order was randomized to avoid learning bias between the URLs.

3.4 Participants

The study was conducted with 89 participants ($N_A = 40$, $N_P = 49$), which were recruited online by posting information about the study in different social network groups of universities and distributing it via university mailing lists. Recruitment advertised the study for people with a general interest in playfully learning about IT security, regular online activities and little to no prior knowledge in IT security and Computer Science. Due to the duration of the study, a financial incentive of 15 EUR was offered to each participant. For participants of the longitudinal testing three months later, a lottery of 4×10 EUR was offered. Both recruiting and financial incentives may have introduced a potential selection bias.

Among the participants, 55.06% identified as female and 44.94% as male. Most participants were between 20 and 29 years of age (76.40%), followed by participants aged 30 or more (16.85%). The analysis of the participants' level of education revealed that most participants were students with either Bachelor's degree or high school diploma (82.02%). Other participants reported to have completed a Master's degree (15.73%), or vocational training (2.25%).

For the longitudinal test three months after the first part of our study, we experienced a dropout of 59.55%, leading to a response rate of only 36 participants ($N_A = 17$, $N_P = 19$). This limits the evaluation of longitudinal effects and calls for reproduction with a larger participant sample.

[2]https://www.alexa.com/topsites/countries online, accessed 2022-02-18

[3]https://tranco-list.eu/ online, accessed 2022-02-18

4 RESULTS

In this section, we attempt to answer the RQs defined in Section 3 using a series of analyses and statistical tests. We first present results of the pre-, post- and longitudinal tests, before analyzing in-game data of the personalized game. For each test, we consider two groups depending on which game the participants played: the *analysis game group* and the *personalized game group*. Note, that longitudinal tests were evaluated only on the reduced set of participants who completed the additional survey.

Performance scores are calculated as the number of correctly classified URLs divided by the total number of URLs was used. Similarly, the confidence levels were computed as the mean confidence of all URLs. Depending on the hypotheses used to answer our research question, one-tailed t-tests or ANOVA were conducted with a significance level $\alpha = .05$. Parametric Student's or Welch's t-tests were used if no deviation from normality was detected in preliminary data screening. Otherwise, non-parametric testing was performed, e.g., Wilcoxon signed-rank test. Effect sizes are provided using either Cohen's d, rank-biserial correlation coefficient r, or partial η_p^2, depending on the computed statistical test.

4.1 Survey Results

Before evaluating our research questions in detail, we check for a potential learning bias on URLs that were present in the pre-test (see Table 3). We therefore compare $M_{\text{post-pre}}$ to $M_{\text{post-new}}$, as well as $M_{\text{long-pre}}$ to $M_{\text{long-new}}$, by performing one-tailed Student's t-tests with the hypothesis that means for URLs that were also used in the pre-test are higher than the new URLs in the post- and longitudinal test. As neither of the two tests is significant ($p > .725$), and means are in fact higher for new URLs in most cases, we argue that learning bias is negligible for our sample.

Next, we analyze the overall effectiveness of the games. Both games were generally effective, in that a one-tailed comparison of pre- and post-test scores (using only URLs that were also present in the pre-test) gives significant results for improvements: Student's t-test for the analysis game with $t_A(39) = 6.404, p_A < .001, d_A = 1.013$ and Wilcoxon signed-rank test for the personalized game with $W_P(48) = 775, p_P < .001, r_P = .717$ (as a deviation from normality was detected; Shapiro-Wilk, $p = .033$).

In response to **RQ-1**, we begin by comparing the post-test results on all 30 post-test URLs of players of the two games, i.e. the analysis game ($N_A = 40$) and the personalized game ($N_P = 49$). Taking a look

Table 3: Means (*M*) and standard deviations (*SD*) for performance and confidence in pre- and post-test including means on partial URL sets for new URLs in post-test (*post-new*) as well as base URLs used in pre- and post-test (*post-pre*).

Game	N	Performance (relative score)				Confidence (range: 1-6)			
		M_{pre} (*SD*)	$M_{\text{post-pre}}$ (*SD*)	M_{post} (*SD*)	$M_{\text{post-new}}$ (*SD*)	M_{pre} (*SD*)	$M_{\text{post-pre}}$ (*SD*)	M_{post} (*SD*)	$M_{\text{post-new}}$ (*SD*)
Analysis	40	.695 (.098)	.828 (.115)	.840 (.095)	.853 (.140)	4.065 (.637)	5.034 (.468)	5.086 (.461)	5.065 (.764)
Personalized	49	.726 (.114)	.811 (.110)	.823 (.104)	.855 (.123)	4.114 (.747)	4.948 (.655)	5.016 (.658)	5.259 (.478)

Table 4: Performance in longitudinal test (*long*), pre- and post-test scores (*pre* and *post-pre*) as well as means of partial URL sets for new URLs in longitudinal test (*long-new*) and base URLs used in pre- and longitudinal test (*long-pre*).

Game	N	M_{pre} (*SD*)	$M_{\text{post-pre}}$ (*SD*)	$M_{\text{long-pre}}$ (*SD*)	$M_{\text{long-new}}$ (*SD*)	M_{long} (*SD*)
Analysis	17	.679 (.095)	.865 (.077)	.812 (.070)	.782 (.119)	.802 (.061)
Personalized	19	.679 (.121)	.800 (.118)	.776 (.112)	.826 (.115)	.793 (.103)

at the mean test results (see Table 3) reveals that personalization did not lead to increased performances or confidences. Even though the analysis game group performed better on average, we did not find this difference to be significant using a two-tailed Welch's t-test ($t(85.891) = .797, p = .428$, $d = .169$, with no deviation from normality: Shapiro-Wilk, $p > .035$). Similar results could be observed for confidence levels: Here, the Shapiro-Wilk test was significant ($p < .001$), a Mann-Whitney test returns no significant results ($U(85.157) = 995.5, p = .901$, $r = .016$).

As it might be possible, that the personalization had an effect on the classification results of different levels of familiarity in the tests, we next perform a repeated-measures ANOVA comparing the three levels of familiarity, with the games as between-groups factor. Note, that $N_A = 34$, $N_P = 39$ in this test, as some participants did not select any services as unknown, known or used. Mauchly's test for sphericity is significant ($p < .001$), and Greenhouse-Geisser corrections are applied ($\varepsilon = .728$). Here, we do not observe significant differences between the two games either: $F(1,71) = .084, p = .772, \eta_p^2 = .001$. We do, however, find significant differences between the levels of familiarity: $F(1.455, 103.308) = 10.204, p < .001, \eta_p^2 = .126$. Post-hoc tests (Holm) confirm, that URLs of unknown services are classified significantly less accurately than known and used in both games ($p <= .001$ in both cases), with no significant differences between known and used ($p = .525$). In all, our study setup did not yield any significant differences of performance scores and confidence levels between the personalized game group and analysis game group.

For **RQ-2**, we are interested in the long-term effect of the two versions of the learning game. Due to a low response rate for longitudinal testing, participant samples are smaller for both groups (N_A = 17, N_P = 19). Data exploration seems to indicate a decline in performance between post- and longitudinal test, with the pre-test score remaining the lowest (see Table 4).

To test for significance of the mean differences, we perform a repeated-measures ANOVA, using the three tests (pre, post, longitudinal) as repeated measures and the games as between-subject factors. Mauchly's test for sphericity is not significant, and the ANOVA ($F(2,68) = 28.432, p < .001, \eta_p^2 = .455$) confirms, that there are significant differences. Post-hoc tests (Holm) show, that pre-test performance is significantly lower than both post- and longitudinal-test performances ($p < .001$ in both cases), while the differences between post- and longitudinal tests are not significant ($p = .074$).

Table 5: **Behavioral Change Questionnaire** results with item group reliabilities (Cronbach's alpha).

Game	$M_{\text{App}}(SD)$	$M_{\text{Int}}(SD)$	$M_{\text{BC}}(SD)$	$M_{\text{PT}}(SD)$
Analysis	3.509 (1.285)	4.059 (0.966)	3.853 (1.412)	4.176 (0.557)
Personal.	4.071 (1.275)	4.684 (1.121)	4.105 (1.174)	4.368 (1.141)

Finally, we take an exploratory look at the results of the self-reported behavioral changes questionnaire of the longitudinal test (see Table 5). As explained in Section 3.3 we split the items of the behavioral change questionnaire into four constructs: whether lessons from the game were applied after playing (Application), how interested participants are in security-related learning games (Interest), whether participants changed their everyday behavior after playing the games (Behavior Change), and to what extend the participants perceive phishing as a threat (Perceived Threat). As expected of a self-reported measure, where we expect a certain amount of bias, the overall results are rather positive (see Table 5). Comparing the mean values, we can observe minor differences between the two groups in all constructs. In particular, the means of the personalized game group are higher in all four constructs. As for differences between the four constructs, it seems that participants were less likely to have applied the learned knowledge and changed their behavior, as the mean scores are lower than the results for "Interest" and "Perceived Threat". Due to the small sample size, we refrain from further statistical testing, but the observed difference calls for more thorough testing in the future.

4.2 In-game Results

To answer **RQ-3**, we perform an exploratory analysis of the game log data of the personalized game. The personalized game gives more insight into the players' interactions with different services during gameplay, as this information is not available for the original analysis game. Python scripts were used to parse the in-game log data and extract different event sequences, including timing information as well as the outcomes of classification events. In the following, mean values are first computed per player and then analyzed, e.g., as the average of all players.

We start by taking a look at the sorting outcomes and time needed for the classification of URLs of different levels of familiarity (see Table 6). We observe notable differences in relative classification outcomes, with URLs of unknown services being classified with the least accuracy with a mean difference of .068 to known and .083 to used services.

Next, we assess the differences in correct classification outcomes per familiarity level per URL category to gain a better understanding of which categories contribute to this difference. As there is a large number of comparisons for all possible levels of familiarity and categories, we focus on percentages of misclassifications (phishing URLs as benign, or benign as phishing URLs), per familiarity level per URL category present in the game (see Table 7). The table also includes the number of valid (and missing) values per category per familiarity, as some players did not classify any URLs of e.g. Path URLs of unknown services. There are only minor differences between the familiarity levels for the URL categories "Path", "IP", and "Random" (mean differences $<= .02$), which were generally detected very well. URLs of the categories "RegDomain" (mean differences $<= .137$) and "No-Phish" (mean differences $<= .044$) have notable differences, with the highest rates of mistakes for unknown services. The classification accuracy for URLs of the "Subdomain" category, interestingly, is highest for unknown services (mean differences $<= .017$). Note, however, that the large number of possible familiarity and category combinations leads to a higher probability of these differences happening by chance.

In all, the detailed analysis of the personalized game seems to indicate, that URLs of unknown services are classified with less accuracy than URLs with services of the other familiarity levels, i.e. used or known, which can mainly be attributed to the URL categories "RegDomain" and "No-Phish".

Table 6: In-game means and standard deviations.

Familiarity	Correct	Incorrect	Unclassified	Time (sec)
Used	.680 (.170)	.186 (.108)	.133 (.117)	4.13 (1.39)
Known	.665 (.180)	.192 (.142)	.143 (.115)	4.11 (1.69)
Unknown	.597 (.221)	.250 (.187)	.154 (.187)	4.27 (1.62)

Table 7: Mean of misclassifications per type per familiarity.

Category	Familiarity	N (Missing)	Mean
IP	unknown	44(5)	.011
	known	48(1)	.006
	used	48(1)	.019
No-Phish	unknown	46(3)	.221
	known	47(2)	.178
	used	49(0)	.177
Path	unknown	27(22)	.000
	known	24(25)	.000
	used	35(14)	.000
Random	unknown	47(2)	.022
	known	49(0)	.008
	used	49(0)	.002
RegDomain	unknown	40(9)	.246
	known	44(5)	.109
	used	44(5)	.153
Subdomain	unknown	40(9)	.096
	known	40(9)	.113
	used	39(10)	.109

5 DISCUSSION

In the previous section, the results of our user study and in-game analysis were described in response to the RQs presented in Section 3. While there were no significant differences in the participants' performance and confidence between the two games (**RQ-1**), we found significant differences between the levels of familiarity with services. In particular, URLs of unknown services were classified significantly less accurately than those of known and used services. For **RQ-2**, longitudinal testing revealed an overall improvement of the participants' performance, since performance means of both post and longitudinal tests are significantly higher than the participants' pre-test performance. Differences based on levels of familiarity were also confirmed in the in-game log analysis in **RQ-3**. In the following, we discuss issues and open questions regarding the overall setup and results of our user study and analysis of in-game behavior.

5.1 Study Setup

Our study setup uses a pre-/post and longitudinal between-group design comparing two versions of the anti-phishing learning game "All sorts of Phish" (Roepke et al., 2021a). Participation for the pre- and post-test was independent from the longitudinal test, which led to high dropout rate of 59.55% and only 36 participants (compared to 89 participants at first). The question arises whether only already interested participants agreed to take part in the longitudinal test, which introduces additional bias, in par-

ticular to the results of the behavioral change questionnaire. As such, results cannot be generalized and we recommend repeating the study with larger sample to strengthen the evidence base. Furthermore, we would be interested in evaluating with even longer time spans to see how the participants' performance changes and whether regular repetitions might be needed in order for the knowledge to remain present.

As the selected game only focuses on teaching essential knowledge about URLs and possible manipulation techniques used for phishing, a limitation of the game as well as the complete study is that we can not make any assumptions on the participants' overall awareness and real-world performance in regards to phishing attacks. Here, we do not claim that the game or its personalized version raise situational awareness and help avoiding phishing attacks in real-world settings. For this, we would recommend additional educational resources to teach how and when phishers lure potential victims into disclosing personal information or redesigning the game to include necessary information and approaches to raise awareness. Whether personalization has an effect on awareness might be an interesting question for future work.

5.2 Study Results

As described in Section 4, we found that while there are no differences between the personalized and non-personalized versions of the games, familiarity with a service did have an effect on the classification outcome in our study (**RQ-1**). We note, that the results of our comparison do not mean that personalization does not have an effect at all, as the URLs that appear in the analysis game were customized and selected to have a high chance of being known by participants. As such, the only difference between the two versions that we can be sure of is the inclusion of the service familiarity selection interface in the personalized game. In particular, it is possible that fixing the ratio of unknown services in the game to different values (currently 20%), or integrating explicit instructions to deal with URLs of unknown services might have an impact on the learning outcome or awareness.

When analyzing the longitudinal test, we found that while the mean performance scores decreased compared to the post-test immediately after playing the game, the scores were still higher than the pre-test (**RQ-2**). Even though the sample size was small, we found significant differences between pre- and longitudinal tests, which implies that the knowledge conveyed in the games was retained, at least partly, by the participants of the longitudinal test. In the self-reported behavioral change questionnaire, we found that players of the personalized game had higher mean values than players of the non-personalized analysis game. We note, however, that these results rely on self-reported data from a custom questionnaire, designed to be used in this study setup. Thus, our findings should only be seen as a first indicator that there might be differences when including personalization in the games, but is far from conclusive evidence. It is possible that personalization makes the game more appealing and its learning content more transferable to the real-world contexts in which users have to deal with potential phishing attacks from services they know and use. Future work might explore how simply making personalization options more present might already lead to a more immersive or relevant gaming experience. In addition, we suggest evaluating the used questionnaire on a larger sample size and with domain experts to strengthen its quality and suitability for future studies.

For **RQ-3**, the analysis of in-game data of the personalized game showed, that there are some URL categories with a larger difference in accuracy when classifying URLs of unknown services. Though we argue that it makes sense that the "RegDomain" and "No-Phish" categories have a high impact, as these URLs can be ambiguous if the original domain is unknown, we also note the interesting finding that the classification of URLs in the "subdomain" category was performed with a higher accuracy for URLs of unknown services. As the difference for the "subdomain" category is small compared to the other categories, it is, however, also possible that the difference is due to chance. A general problem with the analysis of in-game data is, that players might have different strategies when playing the game, e.g., first opening a large number of coins and only classifying the easiest ones. These strategies might have affected the analysis outcomes, in particular some differences might have been inflated by a small number of players.

In all, we found that service familiarity has several effects on the participants' classification abilities. While our study setup and the current version of the games did not exhaust more methods for personalization, we argue that content personalization, which has not been explored in much detail in other domains either, is a worthwhile pursuit. Future work opportunities include the redesign of the personalized game to support adaptive gameplay in which players' actions guide the continuation of the game, as well as the inclusion of contextual information in the games and researching the effect on situational awareness, in particular in a personalized game that closely reflects the players' real-world environments. Further future work lies in the reproduction of our results

with larger participant samples and possibly lower dropout rates in longitudinal testing to strengthen the evidence when answering questions regarding long-term effects.

6 CONCLUSION

In this paper, we present the results of a comparative user study of an anti-phishing learning game and its personalized version as well as an analysis of in-game behavior to understand how personalization influences the participants' gameplay and performance. We find, that users interact differently when confronted with URLs based on services they are not familiar with, both during gameplay and in the URL tests of our user study. While we did not find significant differences in the classification performance of participants of the personalized and non-personalized versions of the game, we find some indications that personalization might potentially have positive effects on the players' awareness. Our work therefore motivates further analyses of learning games with personalized content and how it affects players during and after playing the game. Furthermore, we performed longitudinal testing three months after the game was played and find, that while the participants' performance seems to drop compared to the post-test, it is still significantly higher than the pre-test. These results indicate, that general knowledge about the URL structure and possible manipulation techniques can help users detect malicious URLs even several months after the intervention.

ACKNOWLEDGEMENTS

This research was supported by the research training group "Human Centered Systems Security" sponsored by the state of North Rhine-Westphalia.

REFERENCES

Aleroud, A. and Zhou, L. (2017). Phishing environments, techniques, and countermeasures: A survey. *Computers & Security*, 68:160–196.

APWG (2021). APWG Phishing Activity Trends Report, 3rd Quarter 2021. Technical report, Anti-Phishing Working Group.

Bull, S. (2004). Supporting learning with open learner models. *Planning*, 29(14):1.

Canova, G., Volkamer, M., Bergmann, C., and Reinheimer, B. (2015). NoPhish app evaluation: Lab and retention study. In *NDSS Workshop on Usable Security 2015*, USEC '15, San Diego, California. Internet Society.

Dey, R. and Konert, J. (2016). Content Generation for Serious Games. In Dörner, R., Göbel, S., Kickmeier-Rust, M., Masuch, M., and Zweig, K., editors, *Entertainment Computing and Serious Games: International GI-Dagstuhl Seminar 15283, Revised Selected Papers*, pages 174–188. Springer, Cham.

Drury, V., Roepke, R., Schroeder, U., and Meyer, U. (2022). Analyzing and Creating Malicious URLs: A Comparative Study on Anti-Phishing Learning Games. In *Usable Security and Privacy Symposium 2022*, USEC '22, pages 1–13, San Diego, USA. IEEE. [in publication].

Kaspersky (2021). Spam and phishing in Q3 2021. Technical report, Kaspersky.

Kickmeier-Rust, M. D. and Albert, D. (2010). Micro-adaptivity: Protecting immersion in didactically adaptive digital educational games. *Journal of Computer Assisted Learning*, 26(2):95–105.

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., and Hong, J. (2008). Lessons from a real world evaluation of anti-phishing training. In *2008 eCrime Researchers Summit*, pages 1–12.

Lastdrager, E. E. (2014). Achieving a consensual definition of phishing based on a systematic review of the literature. *Crime Science*, 3(1):1–10.

Law, E. L.-C. and Rust-Kickmeier, M. (2008). 80Days: Immersive Digital Educational Games with Adaptive Storytelling. In *Proceedings of the 1st International Workshop on Story-Telling and Educational Games*, STEG '08, pages 56–62, Maastricht, Netherlands. CEUR.

Roepke, R., Drury, V., Meyer, U., and Schroeder, U. (2021a). Exploring Different Game Mechanics for Anti-Phishing Learning Games. In *Games and Learning Alliance*, GaLA '21, Cham. Springer.

Roepke, R., Drury, V., Schroeder, U., and Meyer, U. (2021b). A Modular Architecture for Personalized Learning Content in Anti-Phishing Learning Games. In *Software Engineering 2021 Satellite Events*, SE-SE '21, Braunschweig, Germany. CEUR.

Roepke, R., Koehler, K., Drury, V., Schroeder, U., Wolf, M. R., and Meyer, U. (2020a). A Pond Full of Phishing Games - Analysis of Learning Games for Anti-Phishing Education. In Hatzivasilis, G. and Ioannidis, S., editors, *Model-Driven Simulation and Training Environments for Cybersecurity*, Lecture Notes in Computer Science, pages 41–60, Cham. Springer.

Roepke, R., Schroeder, U., Drury, V., and Meyer, U. (2020b). Towards Personalized Game-Based Learning in Anti-Phishing Education. In *20th International Conference on Advanced Learning Technologies*, ICALT '20, pages 65–66, Tartu, Estonia. IEEE.

Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L. F., Hong, J., and Nunge, E. (2007). Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, SOUPS '07, pages 88–99, New York, USA. ACM.