Analyzing and Creating Malicious URLs: A Comparative Study on Anti-Phishing Learning Games

Vincent Drury*†, Rene Roepke*‡, Ulrik Schroeder‡, Ulrike Meyer†
† *IT-Security Research Group, RWTH Aachen University* Aachen, Germany
Email: {drury, meyer}@itsec.rwth-aachen.de
‡ *Learning Technologies Research Group, RWTH Aachen University* Aachen, Germany
Email: {roepke, schroeder}@cs.rwth-aachen.de

***Abstract*—Anti-phishing learning games are a promising approach to educate the general population about phishing, as they offer a scalable, motivational, and engaging environment for active learning. Existing games have been criticized for their limited game mechanics, which mostly require binary decisions to advance in the games, and for failing to consider the users' familiarity with online services presented in the game. In this paper, we present the evaluation of two novel game prototypes that incorporate more complex game mechanics. The first game requires the classification of URLs into several different categories, thus giving additional insights into the player's decision, while the second game addresses a different cognitive process by requiring the creation of new URLs. We compare the games with each other and with a baseline game which uses binary decisions similar to existing games. A user study with 133 participants shows, that while all three games lead to performance increases, none of the proposed game mechanics offer significant improvements over the baseline. However, we show that the analysis of the new games offers valuable insights into the players' behavior and problems while playing the games, in particular with regards to different categories of phishing URLs. Furthermore, the user study shows that the participants were significantly better in classifying URLs of services they know than those they do not know. These results indicate, that the distinction between known and unknown services in phishing tests is important to gain a better understanding of the test results, and should be considered when designing and reproducing studies.**

I. Introduction

Phishing attacks still pose a large risk to Internet users world-wide, with more than 700 000 unique websites reported to the APWG in the third quarter of 2021 [3], and more than 45 000 000 clicks on phishing links detected by Kaspersky in Q3 2021 [10]. To address this threat, anti-phishing games have been developed as a scalable and motivating alternative to traditional educational approaches. A common topic of these games is the classification of URLs, as it provides a robust way to determine the origin of a website and can be generalized to several other contexts (e.g., email sender identification). Accurately classifying URLs as phishing or benign requires parsing the URL into different parts and analyzing these parts to understand the destination of the URL. Most existing games use a simple binary (benign/malicious) decision for URL classification during gameplay, which does not reflect this parsing process [21]. Furthermore, existing studies do not consider whether participants are familiar with the services they encounter during the games or the accompanying tests, which could be an explanatory factor for differences when comparing the results of user studies.

In this paper, we address these open questions by presenting a user study that compares three anti-phishing games that mainly differ in their game mechanics (see Section III). The study focuses on the knowledge that is required to detect phishing URLs by analyzing the effect of game mechanics, familiarity with a service, and the structure of phishing URLs on the classification outcome. To this end, two of the proposed games incorporate more complex decision making processes into their game mechanics: The *analysis game* requires players to parse URLs into parts and to sort phishing URLs into several categories based on these parts, while the *creation game* asks players to generate completely new URLs. As the analysis game extends the binary decision used in existing games, we also created the *decision game* as a baseline, which relies on a binary decision scheme and thus enables a comparison based only on the game mechanics (see Section V). Existing games were unsuitable for comparison with our developed game prototypes since they were either unavailable as open source for adaptation or not in the language spoken in our country of origin. While the proposed games focus on the URL structure and possible manipulation techniques and thus, only present a very specific part of anti-phishing education, they could provide one building block of a more comprehensive approach including further educational resources.

We tested the three games in a user study with 133 participants consisting mostly of students between 20-29 years of age. The study setup used between-group comparisons for the three games in a pre-test/post-test design (see Section VI). The main task in pre- and post-test required participants to classify URLs as either benign or malicious and rate the confidence in their decisions. As a result of the user study (see Section VII), we find that the structured approach to URL parsing taught in all three games significantly improves the

* These authors contributed equally.

Usable Security and Privacy (USEC) Symposium 2022
28 April 2022, San Diego, CA, USA
ISBN 1-891562-79-7
https://dx.doi.org/10.14722/usec.2022.23085
www.ndss-symposium.org, https://www.usablesecurity.net/USEC/

classification performance of the participants. While we do not find, that the proposed game mechanics offer significant advantages in the post-test classification task, we show that the game mechanics in the analysis game enable a more detailed analysis of in-game data, including insights into mistakes and possible misconceptions.

The study setup is also designed to validate, whether participants are familiar with services that appear in the games and tests, thus enabling the evaluation of the effect of service familiarity on the classification outcome. We find, that known services are classified significantly more accurately than unknown services, raising questions about the validity of previous surveys where information about participants' familiarity with services is not available. In particular, the familiarity with services should be considered when reproducing existing studies, as it might lead to differences when the classification tasks are adapted to a new population.

Additionally, the example URLs that were selected for the classification tasks in pre- and post-tests are based on eleven URL categories, which are separated by the manipulation method they use (explained in Section IV). Utilizing this categorization, we find that there are categories of phishing URLs that seem particularly hard to detect by users, even after playing the games. Interestingly, commonly used URL highlighting techniques or additionally inspecting domain names in TLS certificates will not make the detection of these URLs easier. The fact that participants were unable to detect these URLs in a lab study might indicate, that even trained users cannot be expected to be able to recognize these phishing URLs in a real-world setting.

II. Related Work

Phishing research encompasses a large amount of work aiming to fight the threat from different, often interacting directions. The most commonly used technical approach uses blocklists of known phishing URLs, which are directly integrated into many popular browsers [13]. While these lists offer low false positive rates and high explainability, the manual submission and review process required for their creation and maintenance leaves victims vulnerable for a short amount of time [24]. Since most phishing websites are short-lived, this short "window of opportunity" can, however, still have severe effects [15]. One way to shorten this window of opportunity is to use machine learning-based approaches that are able to classify new websites on demand (cf. [6]). However, machine learning-based approaches are not as commonly available as the blocklists integrated into browsers. A further technical measure focuses on strong authentication, e.g., Universal Second Factor devices prevent many types of phishing attacks, but are not widely deployed or accepted by users [7].

As a complementary approach to the technical measures, researchers have also turned to studying the human factor in phishing attacks. Here, the question arises why users are susceptible to phishing in the first place. In 2006, Dhamija et al. presented a user study, where they showed 20 websites to their participants and asked them to decide whether the website was legitimate or spoofed [8]. They found, that few users actually made use of indicators in the browser, instead they put more focus on website content. Other studies also include psychological aspects of phishing perception, e.g., using the principles of persuasion [5]. Oliveira et al. studied the differences of susceptibility between younger and older participants and found that older people were generally more likely to click on links in phishing emails, and that the categories had different click-rates for the different age groups [16]. As for the understanding of URL structures, a recent study by Reynolds et al. measured the URL reading ability of users and found, that long subdomains are the most confusing category of URLs, while typo-squatting URLs were recognized with the highest accuracy [18]. These user studies reveal two requirements for users to prevent phishing: they need to possess the knowledge what to look for to detect attacks, and the situational awareness to apply it at the right moment. The study in this paper focuses on the knowledge aspect, and aims to answer how URL knowledge can be applied to detect phishing websites.

To cover these inherent vulnerabilities by teaching the necessary knowledge and raising awareness, researchers have turned to educational approaches. Of particular interest to this paper are educational approaches using game-based learning to improve learning and foster motivation and engagement. Games provide consequence-free environments where learners can experiment and make mistakes, and thus, games allow for graceful failure and active learning [17]. Different topics for anti-phishing learning games are possible, each with potential advantages, as researchers struggle to address the evolving threat. Here, URL classification is a common topic for the games [21], as URLs can not be chosen freely by a phisher and are thus a robust proof of a website's origin, are generally available for phishing attacks that use websites, and are further a common element users encounter when using the Internet. Anti-phishing Phil is an early example of an anti-phishing game, that teaches conceptual and procedural knowledge, followed by levels where players have to classify URLs into benign and phishing URLs [23]. The game was evaluated using a pre- and post-test setup, where 20 websites (ten benign and ten phishing) had to be classified and the confidence about each classification had to be rated. Sheng et al. compared the game to other types of existing educational materials and found, that participants had higher scores and confidences after playing the game. No Phish is a second notable approach to game-based anti-phishing education [4], which requires binary classification (phishing or benign) similar to other games. However it also includes a different type of level where users are instructed to select the registrable domain of URLs, which requires a different cognitive process and makes guessing more difficult. The novel games in this paper extend on this idea, as they also go beyond the binary classification task, with one game even requiring users to create completely new URLs.

A possible problem when reproducing studies of previous work is, that the services used in these studies are often already tailored towards a certain population (e.g., the Bank of America is relatively unknown in Europe). In fact, prior research even hypothesizes that knowledge about services that are presented in games might have an impact on the learning effect of players [22], even though, to our knowledge, this has not been tested yet. To this end, our study includes questions about known and unknown services, and takes a closer look at the differences between distinct levels of familiarity.

To summarize, while most existing anti-phishing games

focus on binary classification tasks, two new games with more complex choices are described in this paper and compared to a binary decision baseline game. Furthermore, existing studies did not consider the question whether the services that were used in the studies were actually known to users, which is a central part of the conducted study.

III. Research Questions and Approach

Based on the identified research gap, we developed three new game prototypes and designed a user study to evaluate them using the following research questions:

- **RQ-1**: Do the games have a positive influence on the participants' performance in classifying URLs?
- **RQ-2**: Are there differences in the participants' performances between the three games? Are there advantages to using the newly proposed game mechanics?
- **RQ-3**: Do participants perform better in classifying URLs of services they know or use?
- **RQ-4**: Are there performance differences in classifying different URL categories?
- **RQ-5**: How is the participants' confidence in classifying URLs influenced by the games?

For the evaluation of our three games and the particular aspects regarding familiarity of services and possible differences between selected URL categories, we designed a user study with three groups of participants, where each group plays one game. The objectives of our research are to evaluate the three games in a pre-/post-test study setup focused on URL classification knowledge. We aim to compare the learning outcomes of all games by comparing the players' performances and confidences after playing either game. We further analyze the effect of familiarity of services on classification performance. Finally, we compare the initial state, improvements and in-game behavior for classifying different URL categories.

IV. URL Categories

In the scope of this work, we differentiate a total of eleven categories of URLs, consisting of one benign and ten malicious categories (see Table I). This categorization enables a more detailed analysis of the results of the URL classification test, as differences between the categories might indicate classes of URLs that are inherently more complicated to detect for users in our study. While the games only teach a simplified version of the categorization to avoid confusion, we can also use the more detailed categories to enable a fairer comparison of the games, as not all categories appear in all games.

The "Benign" category includes all URLs that are considered benign in the context of the study, i.e. all URLs with an existing and benign *registrable domain*[1]. For the categories of phishing URLs, we assume that phishers target a particular original domain and try to make their phishing URLs look as if they belonged to this original domain, such as "ebay-service.com" as a combo-squatting domain of the original "ebay.com". The categories of phishing URLs thus further subdivide the set of possible phishing URLs by the manipulation technique that was used to derive the phishing URL from the original domain. Creating the categories according to manipulation techniques allows us to map phishing URLs to specific parts of the URL structure that include the original domain or a deceptive keyword, which is a requirement for using simplified categories in the games (see Section V-B). We differentiate whether a malicious keyword appears in a *subdomain*, the *registrable domain*, the *path* (including queries and fragments) or not at all. Other URL parts are possible (e.g., authentication information or a port specification in the hostname), but were not included in the tests or games, as they are less common, and in particular did not appear on any benign login page that we encountered. While the "Subdomain" category already includes all phishing URLs with the keyword in the subdomain (including full target embedding [19]), we further divide categories with the keyword in the registrable domain or path. For registrable domain manipulations, we differentiate "Combo-squatting" [11], replacing the TLD ("TLD"), as well as forms of "Typo-squatting" [1]. Note, that we further separate URLs where characters are added ("Addition") or removed ("Omission") from the original domain from other types of typo-squatting domains to enable a more detailed analysis of this category. Similar to previous work [14], we further differentiate URLs with IP addresses as host ("IP-address") from URLs with malicious keywords in the path ("Path"). The "Random" category consists of URLs without any target domain name, thus appearing random and without context. Finally, the category "URL encoding", which obfuscates part of the registrable domain via percent encoding, was added in the tests to see, how participants would react to a manipulation technique that was not part of the games.

TABLE I: Explanation of URL categories and coverage in Analysis (A), Creation (C), and Decision (D) Games

Category/Subcategory*	Explanation	Games
Benign	URLs with unaltered registrable domains	All
IP addresses	Original domain replaced by IP, target in path	A, D
Path	Random domain, target appears in path	All
Random	Domain and path are random, no target appears	A, D
RegDomain	Misleading part included in registrable domain	
Addition	Character added to original domain	A, D
Combo-squatting	Keyword appended to original domain	All
Omission	Character is removed from original domain	A, D
TLD	Original domain, but TLD is replaced	C
Typo-squatting	Character in original domain is replaced/swapped	A, D
Subdomain	Original domain appears as a subdomain	All
URL encoding	Parts of domain are URL encoded	None

*URLs of all categories appear in pre- and post-tests.

Note, that the URL categories in Table I are presented and taught in a simplified way in the games, to avoid confusion and reduce the amount of time that is required (see Section V-B). For example, while the creation game includes a detailed explanation of registrable domains in general, it only includes examples from the "Combo-squatting" and "TLD" categories. The pre- and post-tests, on the other hand, require users to classify URLs from all categories (see Section VI).

The games, as well as the pre- and post-test used in our study, require example URLs, which were selected from a **pool of benign and phishing URLs** that was created as follows. The pool was constructed with the goal to create a representative set of phishing and benign URLs. To this

[1] as defined in https://url.spec.whatwg.org/, online, accessed 2021-11-09

end, popular websites in our country of origin were collected, classified by the type of service they offer to ensure that commonly phished industry sectors are present, and then used to generate the pool of benign and phishing URLs.

In detail, we start by constructing a set of relevant domain names by selecting services of various "types" (e.g., shopping). To this end, the 50 most popular websites in our country of origin were selected (according to Alexa[2]). Through manual review, we removed 20 websites as they are either adult websites or websites whose landing page was not displayed in the language spoken in our country of origin or English. The service names of the remaining websites were extracted and categorized by the type of service the website offers. These types of service were then compared to the most commonly targeted industries according to the APWG [3] and the 10 most commonly phished targets in Phishtank[3] (as determined from more than 250 000 entries). Service types that were included among common phishing targets but not the Alexa list were added by choosing the highest-ranking websites from the Tranco[4] list which fit the type of service and our country of origin. In all, this results in 38 service names and their corresponding registrable domains, which we expect to be relatively well known in our country of origin. The services are further extended by a URL that points to a login form, as determined by manually visiting the website of each service.

The services are then used to generate the URLs that appear in the three games as well as the pre- and post-test. We automatically generate a pool of benign URLs and one pool each for the categories of malicious URLs (e.g., "ebay-service.com" - "Combo-squatting", "pvyq5h4bmj.com/qgxfcvpacj" - "Random") from the set of service names and registrable domains. This automatic generation is based on simple rule-based modifications of the input URL, and also results in "Benign" URLs that are recognizable by their benign registrable domain but might not actually exist in the real world. We then select a set of URLs for the pre- and post-tests (see Section VI-B) and remove these URLs from the pools. The analysis and decision games randomly select examples from the remaining URLs and present them to players of the games for classification. Since the creation game only requires benign reference domain names, we use the registrable domains of the services there.

To assess differences between URL categories, we created a **URL classification test** which is used in the pre- and post-tests. The test consists of a binary classification task, with URLs selected based on the categories described above. One additional constraint is added, as only URLs of actually existing login pages were selected from the benign URLs. The URLs for the pre- and post-test were selected uniformly at random from the pool of available URLs for each URL category (see Table IX in the Appendix). The pre-test consists of 13 malicious URLs, which were selected by choosing example URLs from all categories, with the two possible manipulations of omission (character or dot after www), subdomain (full target embedding separated with dot or comma), and typo (swapping or replacing characters). The 7 benign URLs were selected by first choosing URLs of differing complexities (e.g., having subdomain) and then randomly selecting URLs to obtain 20 pre-test URLs in total. Ten additional URLs are added in the post-test to test for learning bias and were chosen at random to get to a total of 30 URLs (20 malicious and 10 benign). While the content of the URL classification test was equal for all participants, the order of items in the questionnaire was randomized between participants to reduce the influence of potential learning bias of the test items. We decided to only include URLs in the test, not complete website screenshots, as the games focus on URLs, and previous studies have shown, that users sometimes completely ignore this information when classifying websites (see e.g. [2]). We discuss potential problems of this approach in Section VIII.

[2]https://www.alexa.com/topsites/countries online, accessed 2021-02-16

[3]https://www.phishtank.com/ online, accessed 2021-02-16

[4]https://tranco-list.eu/ online, accessed 2021-02-16

V. Game Prototypes

In order to answer our research questions we developed three learning game prototypes. We started by creating two games with new game mechanics, called "All sorts of Phish" and "A phisher's bag of tricks" [5]. "All sorts of Phish" requires players to classify a given URL into several different buckets (benign, 5 different malicious categories) instead of a binary classification. The game "A phisher's bag of tricks" offers a more constructive approach which involves the application of manipulation techniques to create malicious URLs. As such, we refer to the first game as the *analysis game* and the second game as the *creation game* (see Figure 2 in the Appendix for screenshots). For evaluating a baseline, existing games were neither available as open source or in the language spoken in our country of origin, nor implemented as browser games for desktop devices. Thus, they were not adaptable to be compared in our study and we implemented our own baseline, which is an almost exact clone of the analysis game, differing only in the main game mechanic, which is changed to a binary decision scheme. We refer to the baseline game as the *decision game*.

A. Learning Goals

Throughout each of the three games, players learn about the structure of a URL and the method of URL parsing, i.e. reading a URL and identifying the different components. Players are then introduced to a set of manipulation techniques showcasing how benign URLs can be manipulated to become malicious URLs that still look trustworthy.

While the overall learning goal is for end-users to check the URL and classify it as either benign or phishing before clicking on it or before submitting sensitive data on a website, more fine-grained learning goals are defined for the three games and matched with the cognitive process categories in Bloom's Revised Taxonomy (BRT) [12]. Table X in the Appendix provides a complete overview of all learning goals and the mapping to each game. The learning goals of the games overlap when it comes to factual and conceptual knowledge in the lower-order cognitive processes, i.e. *remember* and *understand* in BRT. However, the goals explicitly differ for higher-order cognitive processes: In the analysis game, the learning goals address the cognitive processes of *analyze* and *evaluate*, while the goals in the creation game are focused on *apply* and *create*. The learning goals of the decision game and the analysis game are identical.

[5]Both games are available at https://erbse.elearn.rwth-aachen.de/en/

Since we used BRT as a design aid for game development, the current prototypes focus only on knowledge about phishing, and do not incorporate situational awareness (further discussed in Section VIII). Therefore, the games are limited to URL-based phishing and might not make up a comprehensive phishing education without additional information.

B. Game Content

All three games consist of tutorials to impart knowledge, followed by levels that challenge the players' understanding of the topic. In the tutorials, each part of the URL structure is introduced together with illustrative phishing URLs, that contain suspicious keywords in the corresponding part of the URL. The games teach the general structure of URLs, with a focus on the three main URL parts *subdomains*, *registrable domain*, and *path*. After playing the game, players should therefore be able to identify the registrable domain, analyze it for occurring manipulation techniques, and base their classification decision on the outcome of this process. Even though the URL categories described in Section IV served as inspiration for the games' contents, not all categories are included in all of the game. Preliminary testing showed that the creation game in particular takes a long time to complete and thus, not all categories were included (see Table I for a mapping of categories to games). Meanwhile, the analysis and decision games use a simplified version of the presented URL categorization, which groups all registrable domain manipulations into a single category. This results in the following categories: "No-Phish" (for benign URLs), "IP", "Random", "Subdomain", "Registrable Domain", and "Path", which are introduced successively as players advance in the game.

C. Game Design

The objective of the analysis game is for players to classify multiple URLs by sorting them into the different URL categories. While the tutorials successively introduce the URL categories defined above, levels allow players to practice their knowledge. A level in the analysis game is time-bound (using 60-seconds timer) and challenges players to beat a adaptable score in order to advance in the game. The adaptable timer and score allow for difficulty adaptation, which is not yet exploited in the scope of this study. In the level, players are presented with multiple URLs (depicted as draggable coins which flip upon click and reveal a URL) and a set of different buckets, each representing a specific URL category. The game offers a bucket for benign URLs labeled "No-Phish" as well as one bucket for each already introduced phishing URL category (e.g. "Random", and "IP address" in the first level, depicted in Figure 2 in the Appendix. An additional bucket labeled "No idea" is provided allowing players to discard URLs they are not able to classify confidently. Players can sort URLs into categories by dragging coins or URLs into buckets. When a coin is dropped into a bucket, players receive immediate feedback about the classification result via a colored aura over the bucket and their scores are updated, i.e. increased for correct decisions and decreased for incorrect decisions (discarding URLs does not change the score). The level is finished when the timer runs out. Next, feedback is presented by a review of exemplary correct and incorrect decision the players made during the level. For each incorrect decision, feedback regarding the correct URL category is given. The decision game is structured equivalently, but instead of providing multiple buckets for different phishing URL categories, only one bucket labeled "Phishing" is available.

The objective of the creation game is for players to apply manipulation techniques to create their own malicious URLs. A level in the creation game consists of two to three different tasks called "presets". Each preset poses a challenge for players to apply a specific manipulation technique to a given service (e.g., "ebay.com") and create a malicious but syntactically valid URL (e.g., "ebay.com-signin.ml"). Players are given a set of URL parts, e.g., single characters like "." or "/" but also strings like ".com" or "signin". To complete a given task, players have to drag different URL parts into an initially empty URL bar while making sure, that the created URL follows a valid structure. In addition to the pre-defined URL parts, players can create custom URL parts or even complete URLs using a text input field and the "Generate" button. For submission, players have to click on the button labeled "Verify". Then, a set of automated checks is performed, that test whether the URL is syntactically correct and fits the requirements of the task, and players receive feedback on the successful and failed checks in a pop-up window. If any check fails, players have to revise their created URL and resubmit it. Compared to the analysis game, levels are not time-bound and players have unlimited attempts. For further details regarding the game design of the two games, we refer to [20].

The new game prototypes were created to test, whether the more complex game mechanics lead to better performance when classifying URLs compared to existing games using a binary decision scheme. This binary decision scheme does not allow for fine-grained assessment and feedback, and has a higher probability of guessing correctly. The aim of the more fine-grained assessment is to reduce the probability of guessing, as the number of possible solutions is higher, while also making the analysis of players' in-game data more powerful, as it is possible to better interpret choices regarding the categories of URLs. Furthermore, the URL parsing that is expected in the analysis game reflects a structured approach, which has been shown to be beneficial in identifying the actual target of a URL [18] and might facilitate the detection of phishing URLs. For the design of the creation game, we noticed that none of the existing games allowed for the creation of phishing URLs by manipulating benign URLs, which is supported by an overview of existing games based on a literature review [21]. Although users are not supposed to construct malicious URLs in a real-world scenario, the knowledge on manipulation techniques and the URL structure could be useful in recognizing malicious URLs. Furthermore, the user's active role in the learning process may lead to a deeper understanding compared to the other propose approaches.

VI. User Study

In order to gain insights into the effectiveness of the games and the different game mechanics, a user study was conducted which is presented in the following. The study uses a three-group pre-test/post-test design with A/B testing, a type of between-group design with three experimental groups and no control group. The three games (see Section V) serve as independent variables and participants were assigned randomly

to play one of the games. The performance and confidence in pre- and post-test serve as dependent variables.

A. Participants

The study was conducted in two parts, with 88 participants in November 2020 who played the analysis and creation games, and 45 participants in May 2021 who played the decision game. Recruiting was done online by posting information about the study in different social network groups of universities as well as distributing it via university mailing lists. Recruitment was focused on 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. Since the study required active participation for 60-70 minutes, a financial incentive of approximately 18 USD was offered to each participant. Among the participants, 85 identified as female (63.91%) and 48 as male (36.09%). The majority of participants was between 20-29 years old (78.20%). Due to the methods of recruiting, the majority of participants were students, with a high number of participants reporting their highest degree to be either a Bachelor's degree or high school diploma (81.95%). The remaining participants had mainly either completed their studies with a Master's degree (12.03%) or completed vocational training (3.01%). Besides the 133 participants, an additional five participants were excluded for different reasons: one participant was excluded due to an unrealistic completion time, and four participants had to be excluded due to technical problems during the online survey.

B. Apparatus and Materials

The study was conducted as a remote, online lab study using a video conferencing software and a web browser. For the pre- and post-test phase, an online survey containing the following questionnaires and tests was used:

- **URL classification test**: This test measures the performance and confidence in classifying a set of URLs (see Section IV). For each URL, participants had to decide whether it was benign or phishing, as well as rate their confidence in the decision on a 6-point Likert scale (from 1 = "very uncertain" to 6 = "very certain"). It was utilized in both pre- and post-test with the aim of answering **RQ-1** to **RQ-5**. A list of the used URLs can be found in Table IX in the Appendix.
- **Recognition of Services**: This questionnaire contains a list of services that were targets in the URLs of the URL classification test and participants were asked whether they use the service, do not use but know the service or do not know the service (in response to **RQ-3**). It was used in the post-test and covers services of pre- and post-test URLs (see Table VII in the Appendix for the complete list of services).
- **Demographics**: This questionnaire contains questions regarding gender, age, educational background, Computer Science education and self-reporting of prior knowledge in Computer Science, IT-Security and Phishing. It was used in the post-test (see Table VIII in the Appendix) and was included to report on potential biases among the participants.

C. Procedure

The study began with a briefing phase, where the procedure of the study as well as the requirements for participation were explained. To give more contextual information and establish a shared understanding, a definition of phishing including an example was presented. The decision of which game was to be played by which participant was done uniformly at random by the survey system when each participant started the pre-test phase. Participants were redirected to the game from the survey platform, and did not know that different games were tested in the survey, nor to which group they were randomly assigned. After all participants finished the post-test, the instructors explained the purpose of the study and answered questions of participants in a debriefing before closing the session.

Note, that our institution does not have an ethics committee that could have approved this study. Instead, the study was designed similarly to existing studies with ethical approval. In particular, we were open about the context of the study and the goal of evaluating anti-phishing learning games, and provided additional information as well as a contact email address for participants in case they had questions or concerns after the study. The study also complies with data protection policies as discussed with the data protection officer of our institution, by limiting the collection of identifiable information and replacing names and email addresses of participants with unique random tokens before the analysis.

VII. Results

Based on our research questions described in Section III, we conducted a series of analyses and tests. For each test, we consider three groups depending on which game the participants played: the *creation game group*, the *analysis game group* and the *decision game group*. The performance scores reported in the following are computed as relative scores, i.e. number of correctly classified URLs divided by number of all URLs. The confidence is equal to the average of confidence ratings (ranging from 1 to 6) in either pre- or post-test. Both are measured using an interval scale. We use a significance level $\alpha = .05$. The indices *pre* and *post* are used to distinguish pre- and post-test and hyphenated suffixes in the indices are used to distinguish post-test scores on the URLs also used in the pre-test (*post-pre*) or newly added URLs (*post-new*). Furthermore, indices **A** (for analysis), **C** (for creation) and **D** (for decision) are used to indicate the respective game group.

We first check for a potential learning bias by comparing the performance means in the post-test (see Table III). Instead of a higher mean performance for URLs also used in the pre-test ($M_{\text{post-pre}}$), the mean performance for new URLs in the post-test ($M_{\text{post-new}}$) is higher. As such, we argue that the effect of learning bias is negligible. We therefore compare only those URLs that were part of both pre- and post-test for **RQ-1**, where we look at general improvements, while including all post-test URLs in subsequent sections.

A. Differences Between Pre- and Post-Test

The first RQ described in Section III focuses only on the general effectiveness of the games. For **RQ-1** we derive the following hypothesis: The participants' performance in classifying URLs increased after playing either one of the

games. As shown in Table III, the mean performance score improved in the post-test for all games. To test for significance of these improvements, a one-tailed Student's t-test was performed for each game, comparing the results of the classification task on pre-test URLs between pre- and post-test. A non-parametric Wilcoxon signed-ranked test was performed if a deviation from normality was detected (Shapiro-Wilk test, cut-off value $\alpha < 0.05$, marked with an asterisk). The results (see Table II) indicate, that the participants' performances increased significantly for all three games. There are, however, differences in effect sizes, which are large for the analysis and the decision game, but not for the creation game.

B. Differences Between the Three Games

To address **RQ-2**, we test the following hypothesis: The participants' performance in classifying URLs in the post-test differs between the three games. Mean values in Table III seem to suggest, that players of the creation game performed worse in the post-test than players of the analysis and decision games, who performed similarly well. To test the hypothesis, we compared the performance scores in the post-test of URLs in URL categories that were part of all three games, including post-only URLs, as the higher number of URLs gives a more precise measurement. An ANCOVA was performed, with the games as between group factor, performance in the post-test as dependent variable, and performance in the pre-test as covariate. Levene's test for equality of variances is not significant ($F(2, 130) = 1.207, p = 0.302$). The ANCOVA does not return significant results for the three games as between-subject factor ($F(2, 129) = 0.505, p = 0.605, \eta_p^2 = .008$), only for the pre-test score as covariate ($F(1, 129) = 45.333, p < 0.001, \eta_p^2 = .260$). We therefore retain the null hypothesis, that the differences in post-test performances between the games are not significant. Of particular note is the fact, that the more complex sorting mechanism included in the analysis game did not result in significant differences to the decision game in our study.

C. Differences Between Used, Known and Unknown Services

For the difference between used, known and unknown services (**RQ-3**), we test the following hypothesis: The participants' performance in classifying URLs of services they use or know is better than for services they do not know. Descriptive results seem to indicate significantly higher performance scores for used and known services (see Table IV). To test for significance in performance scores, we performed a factorial repeated-measure ANOVA, with the tests (pre, post) and service familiarity (unknown, known, used) as factors and the games as between-subject factor. As Mauchly's test for sphericity was significant for levels of familiarity ($p < .001$), degrees of freedom were corrected using Greenhouse-Geisser estimates ($\epsilon = .842$). The results of the ANOVA confirm, that familiarity had a significant effect on the performance ($F(1.683, 188.504) = 30.814, p < .001, \eta_p^2 = .216$). Post-hoc tests using Holm's correction confirm significant differences between unknown and known or used services: unknown and known ($p < .001, d = -.654$), unknown and used ($p < .001, d = -6.665$), but not for known and used ($p = .908, d = -.0.11$). None of the interactions (familiarity $\times$ game, familiarity $\times$ test, familiarity $\times$ text $\times$ game) were significant ($p >= .121$). Overall, familiarity with a service has a significant effect on the participants' performance in pre- and post-tests, with significant differences for unknown URLs compared to known and used URLs.

TABLE II: Results of t-tests comparing relative scores in pre- and post-test for all three games

Game	Test statistic	p-value	effect size
Creation	$t(47) = -3.459$	$p < .001$	$d = -0.499$
Analysis	$t(39) = -6.404$	$p < .001$	$d = -1.013$
Decision*	$W = 24.500$	$p < .001$	$r = -0.946$

*Deviation from normality detected.

D. Differences Between URL Categories

For **RQ-4**, we take a look at URL categories, guided by the hypothesis: There are differences in the participants' performance in classifying different URL categories. Table V shows the average scores for our URL categories in the pre-test, as well as the post-test (including post-only URLs) by game. These statistics seem to suggest, that although there are general improvements after playing the games, some categories are less well classified (e.g., TLD, Typo) while others are generally recognized well (e.g., Path, IP). To test for significant differences in performance scores, two repeated-measure ANOVA using the URL categories as repeated-measures factor were performed: the first for URL categories in the pre-test, and the second for URL categories in the post-test with the game as between-group factor. For both tests, Mauchly's test for sphericity was significant ($p < 0.001$), and degrees of freedom were corrected using Greenhouse-Geisser estimates ($\epsilon_{pre} = 0.769, \epsilon_{post} = 0.641$). The first ANOVA returns significant results, indicating that there are already differences in participant's performance for different URL categories in the pre-test ($F(7.687, 1014.641) = 31.266, p < .001, \eta_p^2 = .192$). Post-hoc tests using Holm's correction confirm several significant differences, mainly including the generally well-detected Path URLs and the Typo and Addition URLs, which had very low average detection rates. Similarly, the second ANOVA confirms that differences are still present in the post-test ($F(6.41, 833.238) = 26.757, p < .001, \eta_p^2 = .171$). Post-hoc tests (Holm, averaged over three games) include significant differences for Path (high detection rates), as well as Typo and TLD (low detection rates) URLs. In conclusion, results indicate that some URL categories (e.g., TLD) were significantly more complicated to detect in our tests than others (e.g., Path), in both pre- and post-test.

E. Effects on Participants' Confidence

In response to **RQ-5**, we assessed the participant's confidence when classifying URLs and evaluated possible differences between pre- and post-test and between the games (see Table III). The players' confidences seem to mirror their performances, as they increase after playing the games, with players of the creation game seeming to feel less confident after playing compared to the analysis and decision games. Statistical testing confirms this, as we found significant improvements of confidence between pre- and post-test with a large effect size in all games (see Table VI). For the creation

TABLE III: Means (and standard deviations) for performance and confidence in pre- and post-test including means on partial URL sets

Game	N	performance (relative score)				confidence (range: 1-6)			
		$M_{\text{pre}}\ (SD)$	$M_{\text{post-pre}}\ (SD)$	$M_{\text{post}}\ (SD)$	$M_{\text{post-new}}\ (SD)$	$M_{\text{pre}}\ (SD)$	$M_{\text{post-pre}}\ (SD)$	$M_{\text{post}}\ (SD)$	$M_{\text{post-new}}\ (SD)$
Creation	48	.702 (.122)	.755 (.122)	.782 (.129)	.838 (.163)	4.118 (.720)	4.701 (.625)	4.751 (.642)	4.923 (.723)
Analysis	40	.695 (.098)	.828 (.115)	.840 (.095)	.853 (.140)	4.065 (.637)	5.034 (.468)	5.086 (.461)	5.065 (.764)
Decision	45	.701 (.097)	.818 (.091)	.831 (.097)	.858 (.141)	4.129 (.714)	5.004 (.542)	5.068 (.500)	5.113 (.580)

TABLE IV: Performance scores (and standard deviations) per service familiarity

Game	Used	Known	Unknown
Pre-test	.690 (.192)	.711 (.168)	.561 (.246)
Creation	.809 (.167)	.790 (.140)	.702 (.239)
Analysis	.835 (.148)	.831 (.133)	.720 (.277)
Decision	.858 (.135)	.807 (.140)	.702 (.235)

TABLE V: Mean pre- and post-test relative scores for all URL categories differentiated in the tests

Category	Pre	Post_C	Post_A	Post_D
Benign	0.741	.802	.850	.882
Addition	0.444	.667	.800	.756
Combo	0.707	.833	.675	.700
IP	0.820	.875	.950	1.00
Omission	0.835	.896	.975	.911
Path	0.947	.938	1.00	1.00
Random	0.865	.667	1.00	.978
Subdomain	0.711	.875	.875	.867
TLD	0.609	.792	.700	.578
Typo	0.421	.573	.700	.622
URL encoding	0.895	.906	.912	.900

and decision games we used parametric t-tests but due to a deviation from normality (Shapiro-Wilk test, cut-off value $\alpha < 0.05$, marked with an asterisk in the table) a non-parametric Wilcoxon signed-rank test was used for the analysis game. Next, a one-way ANCOVA was conducted to determine a statistically significant difference between the three games for the participants' confidence in the post-test, on URL categories that were part of all three games, controlling for the participants' confidence in the pre-test. The results indicate significant differences between the games ($F(2, 129) = 5.429$, $p = .005, \eta_p^2 = .078$) after controlling for the confidence in the pre-test, as well as for the covariate of confidence in the pre-test ($F(1, 129) = 79.372, p < .001, \eta_p^2 = .381$). Post-hoc testing using Holm's correction revealed significant differences between creation game and analysis game ($p = 0.015$) as well as creation game and decision game ($p = 0.015$) but not between analysis game and decision game ($p = .892$).

VIII. Discussion and Future Work

In the previous section, we presented the results of our user study to answer the research questions described in Section III. We found, that (**RQ-1**) performances improved significantly after playing either one of the games, (**RQ-2**) there are no differences in performance between the three games, (**RQ-3**) players were significantly better in classifying URLs of services they are familiar with, and (**RQ-4**) there are significant differences in classification performances for different URL categories. Furthermore, we found significant improvements of the participants' confidence after playing either one of the games as well as a significant difference between the creation game and the decision and analysis games (**RQ-5**). In the following, we discuss the setup and results of our study.

TABLE VI: Results of t-tests comparing confidences in pre- and post-test for all three games

Game	Test statistic	p-value	effect size
Creation	$t(47) = -7.850$	$p < .001$	$d = -1.133$
Analysis*	$W = 1.000$	$p < .001$	$r = -0.997$
Decision	$t(44) = -10.273$	$p < .001$	$d = -1.531$

*Deviation from normality detected.

A. Study Setup

For our setup, a general look at the participants of our study reveals a deviation from the general population. Even though we did not recruit participants of a specific age group or occupation, the advertisements for the study were mainly distributed in online social groups for students. As a result, our test population consists mainly of students and does not represent the general population, which might lead to problems in generalizing our findings. In particular, it is possible that these younger people have different states of minds concerning online risks (see e.g. [16]), or have more experience in reading URLs than the general population. However, we argue that the results might be generalized to the population of students between 20-30 years old, which could be substantiated by additional user studies providing supporting evidence. Note, that we did not study the effect of gender and other demographics on the participants' classification performance in detail, as this was not the goal of this study. For gender in particular, we tested its effect on the ANCOVA in the comparison of the three games (**RQ-2**) to eliminate potential biases, and did not find a significant difference between female and male participants. All other statistical tests performed in this study are based on repeated measures. To support further generalization, we suggest replicating the study on a more representative group of participants.

Even though we performed a remote online study, where participants utilized their own, familiar devices, our study design was a lab study, and did not test, how participants would respond to actual phishing attacks in a more realistic setting. The focus of the games is to impart the knowledge required to detect phishing URLs, and the study shows how well this knowledge can be applied in an optimal setting where participants were fully aware of the task. In particular, we do not claim that the games raise situational awareness and help avoiding phishing attacks in a more realistic setting, since

they do not convey the knowledge and awareness of how and when phishers lure potential victims into disclosing personal information. Since we only tested performance scores and confidence levels over a short period of time, a longitudinal study, possibly supplemented by a simulated phishing attack, could provide further results. We also note, that our study was performed during the time of the COVID-19 pandemic, which might have had an impact on participants' state of mind. We did not test for the effect of the pandemic on the participants, and assume that it has not had a more significant effect on our results than other possible limiting but unknown factors.

In the pre- and post-tests of our user study, we asked participants to classify URLs, not screenshots of websites. This is due to the fact, that it has previously been shown that users do not usually look at the URL bar, even in phishing classification tasks (see e.g., [2]). As the focus of the games, and subsequently our study, is knowledge about URLs in general and phishing URLs in particular, and how this knowledge can be utilized to detect phishing websites, we decided to only include URLs in the tests. Furthermore, knowledge about URLs can be applied in several different contexts, as users can analyze URLs before clicking on them, or use the knowledge about domain names to better understand browser URL highlighting, the sender domain in emails, or TLS certificates [9]. As such, we argue that the chosen URL classification task maps the requirements of our study more precisely than a website classification task. It might, however also possibly amplify the effect of unknown services, as sreenshots would offer more context information. However, we argue that the crafted URLs always include a reference to the original service name, and additional information in the website would therefore not have made a significant difference. The only exception are random URLs, which do not include recognizable service names, but were also not included in the evaluation of **RQ-2** and **RQ-3**.

Finally, we note that the tests also include more malicious URLs than benign URLs, which does not realistically reflect the real world situation and might have led to bias in our results. Still, we argue that the improved results for benign URLs (see Table V) in all three games indicate, that players were not choosing "Phishing" more often, and did instead utilize their understanding of URLs to classify URLs more effectively. In addition, confidence improvements might indicate that participants also felt like they were now able to apply their knowledge more effectively, thus deciding more confidently.

B. Study Results

As described in Section VII, we found significant increases in both performance scores and confidence levels from pre- to post-test in all three games (**RQ-1**, **RQ-5**). Taking a closer look at the differences between the three games (**RQ-2**, **RQ-5**), we found that none of the performances for either game significantly deviated from the others, even though players of the creation game were significantly less confident in the post-test. It is noteworthy, that participants who played the creation game usually took more time and asked more questions during the study, as some of them had problems advancing through the game. One possible explanation for the differences is that the requirement to create URLs by themselves posed a higher difficulty and complexity, which resulted in confusion for players who did not really understand the learning content. This might indicate, that the creation mechanic is less suited for self-learning and should be avoided in such contexts. We note, however, that the creation game differs from the analysis (and decision) game in the included URL categories, the tutorials and the levels. As such, the results of this comparison are less significant than the comparison between the analysis and the decision game.

While we did not find significant differences between the analysis game and the decision game, we still argue that the analysis game offers several advantages. In particular, when performing an analysis of in-game data for both games, the analysis game is able to offer more insights into players decision processes. This is due to the fact, that players' mistakes when URLs are sorted into buckets of different URL categories, instead of making a binary decision, offer a better understanding of players' misconceptions. Figure 1 shows the sorting outcomes (i.e. the percentage of URLs sorted into different buckets per URL category, where the outcome "not classified" includes all discarded or opened but not classified URLs) of the analysis game. Even though general trends are visible in the decision game as well, the choices and confusions of players are more evident in the analysis game. In particular, we note that many players had difficulties with Path URLs, often classifying them as random URLs. This trend is not visible in the decision game, as it does not make this distinction, and indicates that players focus mainly on the domain name for classification, while mostly ignoring path information. Furthermore, the analysis game is better suited to understand misconceptions in the basic parsing abilities of URLs. This can, for example, be seen in the case of URLs in the Subdomain category, which were often confused with RegDomain URLs. Since the more complex mechanic also did not lead to drops in performance, we recommend its use if more information on the decision process of players is required in games, as might for example be the case in adaptive games.

For **RQ-3**, we looked at differences between known and unknown services in the post-test, and found significant performance differences in all games. In particular, we found that even though the performance increased for unknown services, they were not at the same level of known or used services, while there was no significant difference between used and known but not used services. Similar results could also be observed for players' confidences, which mirrored the behavior of performance scores. This opens a number of questions regarding the validity of test scores of phishing susceptibility tests in general if services in the test are unknown, as they might no longer accurately reflect the participants' abilities to detect phishing URLs. We note, however, that there were generally less URLs with services that participants did not know than those they did know (3.26 unknown on average), with a few players even knowing all of the services, which might have introduced a bias in the comparison. In the future, we recommend to include some lesser known services in the tests (and the games), to assess potential differences in more detail. We further recommend including a survey for service familiarity when testing phishing classification ability, that at least differentiates used or known from unknown services.

When focusing on different categories of URLs, we found that there are significant differences for performance scores

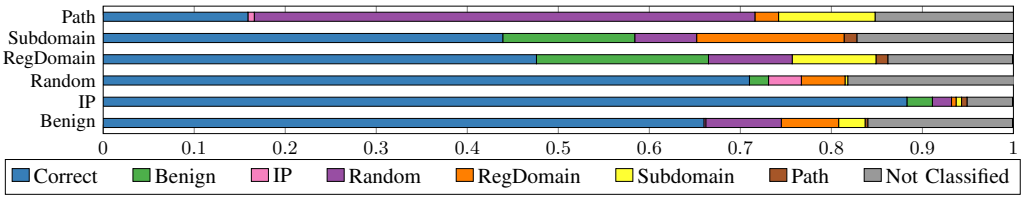


Fig. 1: Relative sorting outcomes for URL categories in the analysis game.

in both pre- and post-test (**RQ-4**). This is consistent with prior work, that showed significant differences in test scores as well [4], [18]. Of particular note are the high scores in URLs that include malicious keywords in the path, especially after playing the analysis or decision games. This seems to indicate, that the distinction between domain name and path is relatively easy to grasp for players of the games, and that phishers might have to create benign-looking domain names to lure educated users. As for subdomains, we saw a significant increase in performance, but participants still seemed to have more trouble recognizing them compared to the "Path" category. The most troubling results are for URLs that manipulate the registrable domain, as these were often detected less accurately, even in the lab setting of our study. This raises the question, whether users can be relied upon to detect these categories of phishing URLs at all, as the URLs cannot be simplified by URL highlighting or looking at domain names in a certificate. We argue that this drawback might be generalized to other educational games and resources as well, since determining the exact registrable domain was a substantial part of all three games, by including conceptual and procedural knowledge in addition to the potential to test this knowledge in the levels. Lastly, URL categories that were not part of the game did not improve to the same extend as included categories, which might imply that knowledge was not transferred and retraining might therefore be necessary for newly emerging categories of phishing URLs. Comparing our results with those in NoPhish [4] or the study by Reynolds et al. [18], pre-test results already differ for our URL categories. These differences might indicate that there is a different measure of difficulty of a phishing URL that is not necessarily connected to the categories used in this paper, or might be due to differences in the study setup. Note, that the number of URLs per category is low in our setup, which might amplify this kind of hidden bias. For more reliable results we would recommend to include more URLs per category in the tests, which might on the other hand increase the average study duration and hence the chance for fatigue among participants.

To summarize, we found that there are differences in improvements of confidence between the three games, however we did not come to a conclusion as to why this is the case. While the differences might indicate, that the creation game is less suited to self-learning in general, future work should test if these differences can be replicated, in particular when the creation game is adjusted to clarify questions that were asked by several of the participants, or even completely aligned with the other two games. Furthermore, we found that participants performed significantly worse when classifying URLs for unknown services in our study setup, which raises questions about the reproducability of studies that do not consider the familiarity of participants with the services used in their tests. We aim to analyze this difference more closely in the future, to find out whether differences can already be observed while playing the games, by utilizing learning analytics approaches on the games' log data. Similarly, differences in classification performances also exist for different URL categories, where manipulation techniques targeting the registrable domain or subdomains led to lower accuracy than those targeting the path. Future work might test, whether different tutorial content or levels can improve the accuracy for these categories. Finally, a test of the longitudinal effects as well as a test in a more realistic setting and with representative demographics, with a more comprehensive set of educational resources, might be insightful additions in future work.

IX. Conclusion

In this paper, we present the results of a user study that tested the participants' performance and confidence when classifying URLs before and after playing one of three anti-phishing learning games. Unlike most previous games, two of the games of this work actively incorporate more complex decision schemes by providing different game mechanics, e.g. for creating URLs. The third game, however, is designed similar to existing games to serve as a baseline. We did not find significant differences between the three games in performance scores, only in confidence levels, indicating that the type of game might have an impact on the players' understanding, and posing the question whether this difference can be explained by a fundamental difference between the games in future studies. Further results indicate, that the familiarity with services that appeared in the tests and games had a significant effect on the performance scores. In particular, services that were known or used by users were recognized better than unknown services. This effect has been ignored by previous games and studies, and raises a number of questions for future work. As such, we aim to further study the difference between known and unknown services, introduce personalization options to anti-phishing learning games and analyze the players' in-game actions in more detail. Since our games only convey knowledge about the URL structure and possible manipulation techniques, we also plan to embed them into a suitable learning context with additional resources to raise situational awareness and address other aspects of phishing.

Acknowledgments

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

REFERENCES

[1] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis, "Seven Months' Worth of Mistakes: A Longitudinal Study of Typosquatting Abuse," in *Network and Distributed System Security Symposium*, ser. NDSS '15. San Diego, CA: Internet Society, 2015.

[2] M. Alsharnouby, F. Alaca, and S. Chiasson, "Why phishing still works: User strategies for combating phishing attacks," *Human-Computer Studies*, vol. 82, pp. 69–82, 2015.

[3] APWG, "APWG Phishing Activity Trends Report, 3rd Quarter 2021," Anti-Phishing Working Group, Tech. Rep., 2021. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q1_2021.pdf

[4] G. Canova, M. Volkamer, C. Bergmann, and B. Reinheimer, "NoPhish app evaluation: Lab and retention study," in *NDSS Workshop on Usable Security 2015*, ser. USEC '15. San Diego, California: Internet Society, 2015.

[5] R. B. Cialdini, *Influence: The Psychology of Persuasion, Revised Edition*. New York: New York: William Morrow, 2006.

[6] A. Das, S. Baki, A. El Aassal, R. Verma, and A. Dunbar, "SoK: A Comprehensive Reexamination of Phishing Research From the Security Perspective," *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 671–708, 2020.

[7] S. Das, A. Dingman, and L. J. Camp, "Why Johnny Doesn't Use Two Factor A Two-Phase Usability Study of the FIDO U2F Security Key," in *Financial Cryptography and Data Security*, S. Meiklejohn and K. Sako, Eds. Berlin: Springer, 2018, pp. 160–179.

[8] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," in *Conference on Human Factors in Computing Systems*, ser. CHI '06. New York: ACM, Apr. 2006, pp. 581–590.

[9] V. Drury and U. Meyer, "Certified phishing: Taking a look at public key certificates of phishing websites," in *Symposium on Usable Privacy and Security*, 2019.

[10] Kaspersky, "Spam and phishing in Q3 2021," Kaspersky, Tech. Rep., 2021. [Online]. Available: https://securelist.com/spam-and-phishing-in-q3-2021/

[11] P. Kintis, N. Miramirkhani, C. Lever, Y. Chen, R. Romero-Gómez, N. Pitropakis, N. Nikiforakis, and M. Antonakakis, "Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse," in *Conference on Computer and Communications Security*, ser. CCS '17. New York: ACM, Oct. 2017, pp. 569–586.

[12] D. R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview," *Theory Into Practice*, vol. 41, no. 4, pp. 212–218, 2002.

[13] A. Oest, Y. Safaei, A. Doupé, G.-J. Ahn, B. Wardman, and K. Tyers, "PhishFarm: A Scalable Framework for Measuring the Effectiveness of Evasion Techniques against Browser Phishing Blacklists," in *Symposium on Security and Privacy (SP)*. IEEE, May 2019, pp. 1344–1361.

[14] A. Oest, Y. Safei, A. Doupé, G.-J. Ahn, B. Wardman, and G. Warner, "Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis," in *APWG Symposium on Electronic Crime Research (eCrime '18)*, May 2018, pp. 1–12.

[15] A. Oest, P. Zhang, B. Wardman, E. Nunes, J. Burgis, A. Zand, K. Thomas, A. Doupé, and G.-J. Ahn, "Sunrise to Sunset: Analyzing the End-to-end Life Cycle and Effectiveness of Phishing Attacks at Scale," in *USENIX Security Symposium (USENIX Security '20)*, 2020, pp. 361–377.

[16] D. Oliveira, H. Rocha, H. Yang, D. Ellis, S. Dommaraju, M. Muradoglu, D. Weir, A. Soliman, T. Lin, and N. 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 *Conference on Human Factors in Computing Systems*, ser. CHI '17. New York: ACM, May 2017, pp. 6412–6424.

[17] J. L. Plass, B. D. Homer, and C. K. Kinzer, "Foundations of Game-Based Learning," *Educational Psychologist*, vol. 50, no. 4, pp. 258–283, 2015.

[18] J. Reynolds, D. Kumar, Z. Ma, R. Subramanian, M. Wu, M. Shelton, J. Mason, E. Stark, and M. Bailey, "Measuring Identity Confusion with Uniform Resource Locators," in *Conference on Human Factors in Computing Systems*, ser. CHI '20. New York: ACM, Apr. 2020, pp. 1–12.

[19] R. Roberts, Y. Goldschlag, R. Walter, T. Chung, A. Mislove, and D. Levin, "You Are Who You Appear to Be: A Longitudinal Study of Domain Impersonation in TLS Certificates," in *Conference on Computer and Communications Security*, ser. CCS '19. New York: ACM, Nov. 2019, pp. 2489–2504.

[20] R. Roepke, V. Drury, U. Meyer, and U. Schroeder, "Exploring Different Game Mechanics for Anti-Phishing Learning Games," in *Games and Learning Alliance*, ser. GaLA '21. Cham: Springer International Publishing, 2021.

[21] R. Roepke, K. Koehler, V. Drury, U. Schroeder, M. R. Wolf, and U. Meyer, "A Pond Full of Phishing Games - Analysis of Learning Games for Anti-Phishing Education," in *Model-Driven Simulation and Training Environments for Cybersecurity*. Cham: Springer International Publishing, 2020, pp. 41–60.

[22] R. Roepke, U. Schroeder, V. Drury, and U. Meyer, "Towards Personalized Game-Based Learning in Anti-Phishing Education," in *International Conference on Advanced Learning Technologies*, ser. ICALT '20. Tartu, Estonia: IEEE, 2020, pp. 65–66.

[23] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. 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*, ser. SOUPS '07. New York: ACM, 2007, pp. 88–99.

[24] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in *Conference on Email and Anti-Spam*, ser. CEAS '09. Mountain View, CA, USA: Carnegie Mellon University, Jul. 2009.

APPENDIX

TABLE VII: Absolute (and relative) results of the **Recognition of Services** questionnaire.

Service	Used	Known	Unknown
Amazon	124 (93.23%)	9 (6.77%)	0
Commerzbank	20 (15.04%)	109 (81.96%)	4 (3.01%)
Deutsche Bank	16 (12.03%)	113 (84.96%)	4 (3.01%)
Dropbox	96 (72.18%)	35 (26.32%)	2 (1.51%)
eBay	90 (67.67%)	43 (32.33%)	0
eBay Kleinanzeigen	106 (79.70%)	27 (20.30%)	0
Facebook	106 (79.70%)	27 (20.30%)	0
FOCUS	16 (12.03%)	105 (78.95%)	12 (9.02%)
GMX	39 (29.32%)	81 (60.90%)	13 (9.77%)
iCloud	53 (39.85%)	75 (56.39%)	5 (3.76%)
ImmobilienScout24	49 (36.84%)	80 (60.15%)	4 (3.01%)
Microsoft	104 (78.20%)	29 (21.81%)	0
Netflix	108 (81.20%)	25 (18.80%)	0
OTTO	42 (31.58%)	87 (65.41%)	4 (3.01%)
PayPal	113 (84.96%)	20 (15.04%)	0
Reddit	27 (20.30%)	71 (53.38%)	35 (26.32%)
Steam	27 (20.30%)	52 (39.10%)	54 (40.60%)
Twitch	19 (14.27%)	77 (57.90%)	37 (27.82%)
VK	3 (2.26%)	23 (17.29%)	107 (80.45%)
WEB.DE	43 (32.33%)	71 (53.38%)	19 (14.29%)
YouTube	121 (90.98%)	12 (9.02%)	0

TABLE VIII: **Demographics** questionnaire including answer types and options

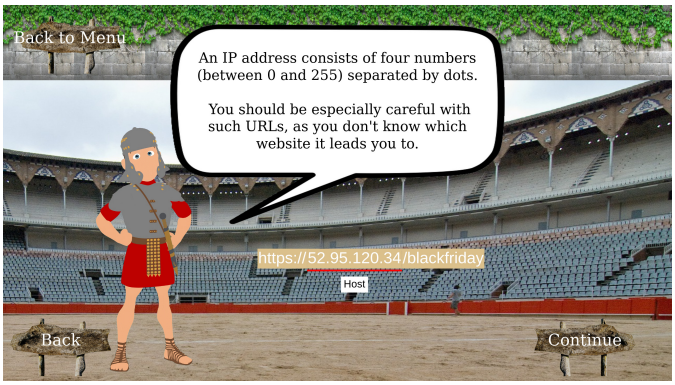
Question	Answer type	Answer options
What is your gender?	single-choice	Female; Male; Diverse; No answer
How old are you?	single-choice	14 or younger; 15-19; 20-24; 25-29; 30-34; 35-39; 40 and older; No answer
What is your highest degree?	single-choice	No school degree; Middle school; High school graduate, diploma or the equivalent; Vocational Training; Bachelor's degree; Master's degree; Diploma; Doctorate degree; Other; No answer
Did you participate in Computer Science classes (e.g. in school or university)?	single-choice	No Computer Science classes; Less than 6 months; 6 to 12 months; 1 to 2 years; More than 2 years; No answer
How would you rate your prior knowledge in the following topics? (Computer Science — IT-Security — Phishing)	6-point Likert scale	None; Very little; Little; Some; Much; Very much

TABLE IX: URLs of the **URL classification test** in pre- and post-test and their mean performance scores (and confidence levels)

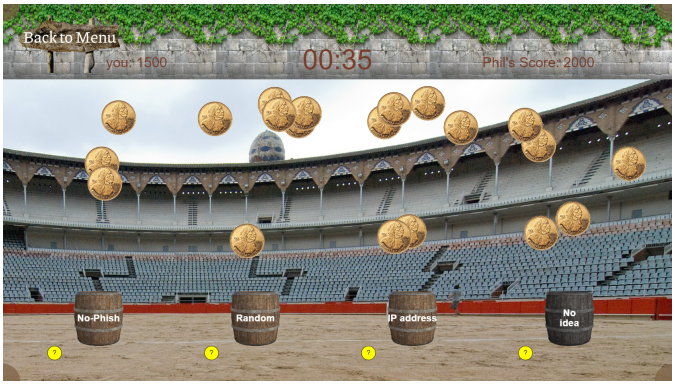
URL	Category	Pre	Post$_C$	Post$_A$	Post$_D$
https://www.otto.de/user/login?entryPoint=loginArea	Benign	.947 (4.481)	.875 (4.792)	.950 (5.300)	1 (5.422)
https://www.amazon.de/ap/signin?openid.pape ...	Benign	.504 (3.729)	.667 (4.313)	.850 (4.800)	.911 (4.867)
https://www.reddit.com/login/	Benign	.962 (4.962)	.958 (5.146)	1 (5.600)	1 (5.644)
https://accounts.google.com/signin ...	Benign	.579 (3.586)	.604 (4.271)	.675 (4.600)	.711 (4.333)
https://meine.deutsche-bank.de/trxm/db/	Benign	.496 (3.789)	.688 (4.375)	.725 (4.475)	.689 (4.578)
https://www.gmx.net/	Benign	.932 (4.842)	.854 (4.979)	.950 (5.500)	.978 (5.444)
https://vk.com/	Benign	.767 (4.000)	.979 (4.833)	.750 (4.600)	.867 (4.800)
https://v-k.com/	Addition	.444 (3.677)	.479 (4.333)	.650 (4.725)	.556 (4.289)
https://amazon-secureserver.de/ap/signin?openid ...	Combo	.707 (3.639)	.896 (4.771)	.800 (4.625)	.911 (4.844)
https://214.156.43.197/login.live.com/	IP	.820 (3.917)	.875 (4.458)	.950 (5.650)	1 (5.644)
https://sso.immobilienscout24.de/sso/login	Omission	.504 (3.955)	.500 (4.729)	.625 (4.625)	.511 (4.667)
https://wwwcommerzbank.de/lp/login	Omission	.835 (4.677)	.896 (5.188)	.975 (5.600)	.911 (5.267)
https://b1ovam5.org/otto.de/	Path	.947 (4.316)	.938 (5.292)	1 (5.425)	1 (5.644)
https://uyvgo8i.net/RsHZdqidvhidpFbRVa/account ...	Random	.865 (4.090)	.583 (4.125)	1 (5.550)	.978 (5.444)
https://www.netflix.com-co.support/login	Subdomain	.632 (4.075)	.833 (4.750)	.850 (4.925)	.800 (4.689)
https://ebay.de.login.9ontzcjkgj2k.ru/ws/eBayISAPI.dll ...	Subdomain	.789 (4.105)	.917 (5.125)	.975 (5.425)	.978 (5.387)
https://meine.deutsche-bank.online/trxm/db/	TLD	.609 (3.774)	.792 (4.458)	.700 (4.500)	.578 (4.089)
https://mircosoft.com/login.srf?wa ...	Typo	.504 (4.188)	.438 (4.813)	.725 (5.375)	.600 (5.533)
https://store.steamposered.com/login/	Typo	.256 (4.000)	.417 (4.354)	.450 (4.375)	.444 (4.444)
https://gmx.net%6B%73%35%66%6C%6A%33%2E ...	URL encoding	.895 (4.308)	.917 (4.917)	.950 (5.000)	.933 (5.067)
https://www.focus.de/ajax/login/community_login ...	Benign	–	.646 (4.208)	.725 (4.675)	.733 (4.778)
https://www.netflix.com/de-en/login	Benign	–	.875 (4.917)	.950 (5.500)	1 (5.689)
https://web.de/	Benign	–	.875 (5.146)	.925 (5.250)	.933 (5.222)
https://www.paypall.de/signin?SignIn&UsingSSL=1& ...	Addition	–	.854 (5.271)	.950 (5.625)	.956 (5.756)
https://www.dropbox-account.com/login?hl=de& ...	Combo	–	.771 (4.479)	.550 (4.725)	.489 (4.400)
https://www.yi19p83.info/iWOaXLrmMRaymXsqdl/ ...	Random	–	.750 (4.438)	1 (5.600)	.978 (5.511)
https://icloud.com-de.support/	Subdomain	–	.875 (4.729)	.725 (4.600)	.756 (4.644)
https://www.twitch.tv.support.i1oc8c8.3pyozv3n ...	Subdomain	–	.875 (4.958)	.950 (5.450)	.933 (5.289)
https://netglix.com/de-en/login	Typo	–	.938 (5.479)	1 (5.725)	.933 (5.822)
https://www.dropbox.com%70%6C%79%74%67 ...	URL encoding	–	.896 (4.896)	.875 (4.750)	.867 (4.844)

TABLE X: Learning goals including the mapping to all three games. Learning goals are marked with x if they apply to the analysis game (A), creation game (C), or decision game (D)

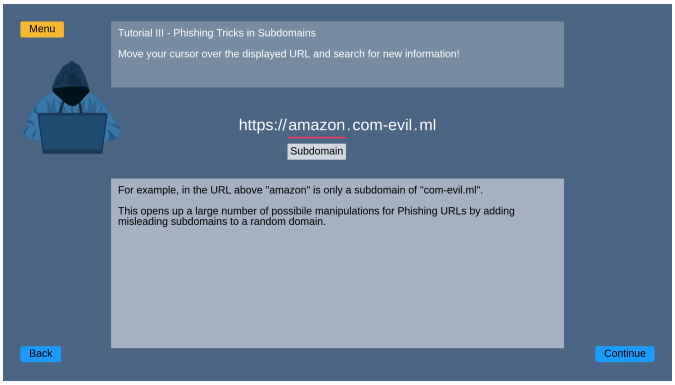
After playing the learning game, players should be able to ...			A	C	D
Remember	...	know the structure of URLs by recalling its components.	x	x	x
	...	name the manipulation techniques for URLs by listing the manipulation techniques for individual components.	x	x	x
	...	know the manipulation techniques for URLs by describing the manipulation of the components.	x	x	x
Understand	...	understand the structure of URLs by explaining the purpose of the components.	x	x	x
	...	understand the manipulation of the structure of URLs by explaining manipulation techniques for the components.	x	x	x
Apply	...	determine the individual components of a URL by performing URL parsing.	x	x	x
	...	compose valid URLs by combining the (necessary) components in the correct order.		x	
	...	compose valid URLs by creating the (necessary) components in the correct order.		x	
	...	change the structure of a URL by modifying components.		x	
	...	manipulate the structure of a URL by modifying (necessary) components based on specific rules.		x	
Analyze	...	analyze the structure of a URL by identifying the components.	x	x	x
	...	detect manipulations in the structure of a URL by identifying manipulated components.	x		x
	...	recognize the manipulation technique applied to a URL by identifying/recognizing the manipulated component.	x		x
Evaluate	...	assess the correctness of the structure of a URL by checking the components.	x		x
	...	assess the manipulation of the structure of a URL by checking the components and identifying manipulated components.	x		x
	...	distinguish benign URLs from manipulated URLs by comparing both URLs in terms of applied manipulation(s).	x		x
Create	...	create correct URLs by creating and combining the (necessary) components.		x	
	...	create manipulated URLs by manipulating and combining (necessary) components based on rules and the URL structure.		x	



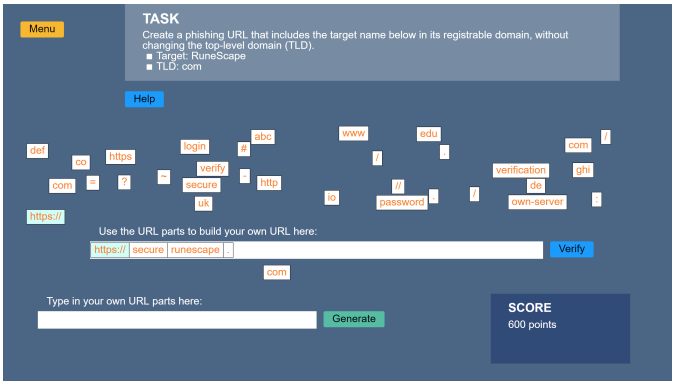
(a) Example tutorial section in the analysis game.



(b) First level of the analysis game. Players have to classify given URLs, which are hidden behind coins.



(c) Example tutorial section in the creation game.



(d) Level of the creation game. Players create malicious URLs by combining URL parts via drag-and-drop.

Fig. 2: Screenshots from the creation and analysis games.