Exploring and Evaluating Different Game Mechanics for 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*
{roepke, schroeder}@cs.rwth-aachen.de
[2] *IT-Security Research Group, RWTH Aachen University*
{drury, meyer}@itsec.rwth-aachen.de

Abstract

Anti-phishing learning games are a promising approach for teaching end-users about phishing, as they offer scalable, engaging learning environments. Existing games have been criticized for limited game mechanics that do not allow detailed assessments of players' acquired knowledge, instead focusing mostly on factual and conceptual knowledge to remember or understand. With the aim of evaluating the effects of new game mechanics on the classification performance of URLs as phishing or benign, and on the understanding of in-game decisions, this paper presents the design and evaluation of two new learning games targeted at end-users who do not necessarily have previous knowledge of IT security: The first game implements extended classification mechanics to better assess players' decision processes, while the second game implements different mechanics, asking players to combine URL parts when creating their own phishing URLs. In a case study with 133 participants, we compared the games with each other and with a third baseline game using binary decisions similar to related work. The study shows, that while all games lead to performance increases, new games do not offer significant improvements over the baseline. Longitudinal tests three months later show that knowledge can be retained as participants still performed significantly better than before playing either of the games.

Keywords: Learning Games, Game Design, Learning Goals, Phishing

1 Introduction

When using the Internet, end-users are immediately exposed to various threats, one of which is phishing, a deception-based threat that uses impersonation to obtain information from a target [1]. Phishing attacks pose a large risk to end-users, with over 700 000 unique websites reported [2] and 45 000 000 clicks on phishing links detected [3] in the third quarter of 2021. Although technical countermeasures exist, they fail to stop the threat completely [2]. As a complementary approach, user education can support end-users in acquiring relevant knowledge and skills to recognize and mitigate phishing. Game-based anti-phishing education is an approach that gained a lot of interest over the last decade. A common topic in anti-phishing learning games is the classification of URLs. It provides a robust way to determine a website's origin and can be generalized to different contexts (e.g., email sender identification). However, recent reviews of anti-phishing learning games [4–6] indicate limited success and distinct commonalities among available games: a focus on factual and conceptual knowledge as well as limited use of game mechanics, as they mostly rely on binary decision mechanics in

which players have to classify emails or URLs as phishing or benign. Even though the binary classification matches the decisions end-users have to make in real life when confronted with a potential phishing URL or email, more complex mechanics might be able to better map the decision processes that lead to the final decision, thus giving insights into the conception of players and potentially facilitating learning the steps of the decision process. In particular, existing games fail to provide detailed assessment of players' decision processes, e.g. misconceptions when making mistakes are not detected due to a high guessing probability and no requirement of providing reasons for decisions. Even beyond the domain of anti-phishing learning games, decisions with binary outcomes are often based on a more complex decision process (e.g., medical diagnoses). Therefore, the results from this paper provide a case study on the potential benefits of extending binary decisions and exploring different game mechanics in learning games in general.

This paper addresses this potential benefit by presenting the design and evaluation of two new game prototypes, which on the one hand extend the use of game mechanics to provide more detailed assessment and feedback, and on the other hand, explore different game mechanics to foster higher-order cognitive processes: The *analysis game* implements extended classification mechanics to better assess players' decision processes, while the *creation game* implements different game mechanics, which requires players to combine URL parts to construct their own phishing URLs. Both games were designed for a target group without necessarily any prior knowledge of IT security and, in particular, phishing. While we did not focus on a specific demographic or age group during the development, game content like instructions and explanations require a basic language proficiency, which is why the games might not be suitable for younger students and children. To provide a comparison to the commonly used binary decision mechanics found in related work, where players only have the (binary) choice between benign and phishing for a given URL, a third game prototype was implemented to serve as a baseline.

In a user study with 133 participants, we evaluated the games in a between-group pre-test/post-test design with additional longitudinal testing. The main task in the study required participants to classify URLs as either benign or malicious and rate their confidence in their decisions. Participants were tested before and after playing one of the games as well as three months later to analyze long-term effects. Results show that the structured approach to URL parsing taught in all three games significantly improved the participants' performance and confidence when classifying URLs. While we did not find that the new games offer significant advantages over the baseline implementation, we show that the game mechanics in the analysis game enables a more detailed analysis of in-game data, including insights into mistakes and possible misconceptions. Furthermore, longitudinal tests show that knowledge is retained, and participants still performed significantly better than in the pre-test. Overall, this work can be seen as a case study exploring how different, extended game mechanics which better reflect players' learning and decision processes can provide better insights and more detailed assessment of gameplay.

2 Related Work

The concepts of serious games and game-based learning have been studied and explored in various application domains, e.g., health, society, science, or business [7]. Argued benefits range from learning-related outcomes, e.g., knowledge and skill acquisition, over perceptual, cognitive, and affective outcomes to physiological outcomes [8]. Over the last decade, various learning games for security education have been developed with the overall goal to foster secure online behavior and provide opportunities to learn and practice. Alongside these de-

velopments, different researchers reviewed existing games to analyze the achievements and challenges within the research domain of security education [5, 6, 9–14]. In these reviews, researchers identified the need for more extensive research [9] and to further explore interactive, immersive simulations for practical, hands-on training [10, 13, 14]. Existing user studies in game-based security education showed positive effects, but they are criticized for small sample sizes and omitting effect sizes [11]. Lastly, although many games are found in literature, the availability of learning games for end-users is limited [12].

While reviews exist that compare games for the broad field of security education, less review work exists focusing solely on anti-phishing learning games. Roepke et al. identified the need for a more in-depth review of anti-phishing learning games in [6], where they created and reviewed a data set on available publications on games. They analyzed publications using Bloom's Revised Taxonomy (BRT) [15] and found that the majority of games convey factual or conceptual knowledge and skills. Other cognitive processes along the BRT and procedural knowledge are mainly neglected by existing games. Lastly, the review of covered learning content in available games showed that the focus often lies on phishing URLs and emails. More complex or advanced contemporary phishing techniques are usually neglected by existing games. The review also revealed that most games rely on very similar game mechanics: a binary decision where players have to decide whether a given URL or email is malicious or benign (e.g., [16, 17]). This type of assessment is very limited since it does not convey how and why players decide. The mechanics also force players to make a decision even if they are unsure, and lastly, feedback is limited since the decision does not reflect possible misconceptions. Only one game provides additional mechanics where players have to select different components of given URLs to show that they can parse URLs [18].

All in all, the limited use of game mechanics results in limited assessment possibilities. While focusing on factual and conceptual knowledge to remember and understand, existing games also miss the opportunity to target other cognitive processes and convey procedural knowledge. Due to missing alignment of learning mechanics and game mechanics, the game design might be insufficient and could be improved. Related work shows that for learning games in general, the coherent mapping between learning mechanics and game mechanics plays a vital role in game design [19, 20] and can be supported using the LM-GM model [21]. The model can be used for game design or for game analysis to understand how learning activities are translated into gameplay and has been applied in various studies (e.g., [22–24]). Although existing games might cover necessary learning content (e.g., about phishing URLs), the gameplay may not support the assessment of players' abilities. To this end, learning and game mechanics may also need to be aligned with assessment mechanics [25]. Conclusively, we identify the potential for new learning games, exploring different game mechanics, and providing more in-depth assessments. The games should also provide gameplay supporting procedural knowledge acquisition and address higher-order cognitive processes.

3 Designing New Game Prototypes

With existing games mainly focusing on factual and conceptual knowledge and being limited in their use of different game mechanics, we decided to explore the design of new anti-phishing learning games incorporating different game mechanics and addressing further cognitive processes of BRT. As such, we formulated two design goals:

1. Extend the binary decision game mechanics of classifying phishing URLs to provide better assessment and feedback
2. Address higher-order cognitive processes of applying, analyzing, evaluating, and creating by using different game mechanics

To evaluate the effects of new game mechanics on classification performance and understanding of in-game decision processes, two novel learning games were implemented fulfilling the design goals. While the first game extends the decision mechanics often used in existing games by providing multiple options to classify a URL, the second game replaces the currently favored classification mechanics by introducing a puzzle or combining game mechanics in which players have to create a solution to a given task by combining URL pieces. Instead of fulfilling both design goals in one game, the decision was made to implement two independent games and later compare them to understand how different game mechanics may affect players' performance and confidence in the scope of a user study. If a combined prototype had been developed and evaluated in a comparative user study, possible interaction effects of the different game mechanics could influence the results. Thus, a direct comparison of different game mechanics would not be possible.

3.1 Learning Content

The main requirement for the games' learning content was to provide knowledge and skills that are simple to understand, retain general applicability, and are robust against adversarial influence. We decided to focus on URLs, as they uniquely identify websites, and are presented by the browser and can therefore not be altered by an attacker. They are also applicable in different types of phishing attacks and can always be used to identify potential phishing websites. As such, the games aim to teach the basics of URLs and, in particular, how to identify the registrable domain of a URL, as it is the discerning factor to decide between legitimate and phishing URLs. The learning content focuses on the structure of the URL[1] with three distinct components: (1) subdomains, (2) registrable domain (RD), and (3) path. The example URLs are based on the login websites of a set of existing services that are popular in our country of origin (e.g., 'eBay' or 'PayPal'; based on Alexa[2] and Tranco[3]).

Furthermore, the games aim to teach a number of manipulation techniques that attackers use to construct phishing URLs, as they can support a better understanding of how to identify potential phishing URLs. These manipulation techniques create a foundation for what to look out for in phishing attacks while strengthening the players' understanding of URLs. They are based on phishing attacks found in the real world and described in related work (see, e.g., [26–28]). Attackers typically construct phishing URLs that make users believe that they lead to a specific benign target domain. We refer to this target domain as the 'original domain'. The manipulation techniques included in the games are:

- `IP address`: Using an IP address as host and (parts of) original domain in path
- `Path`: Using random domain and (parts of) original domain in path
- `Random`: URL is completely random and shows no connection to original domain
- `RD`: Modifying (e.g., adding to, replacing characters in) RD of original domain
- `Subdomain`: Including (parts of) original domain in subdomain
- `Top Level Domain (TLD)`: Using original domain, except for changing TLD

In all, the games aim to impart the necessary knowledge and skills to detect phishing attacks. However, to successfully protect against an actual attack, the user also needs the awareness to apply the conveyed knowledge and skills, which requires a behavioral change in

[1] as defined in `https://url.spec.whatwg.org/`, accessed on 16.02.2022
[2] `https://www.alexa.com/topsites/countries`, accessed on 16.02.2022
[3] `https://tranco-list.eu/`, accessed on 16.02.2022

their daily activities. This would require information about when and where phishing URLs could be encountered and strategies applicable to everyday life, such that users would be able to assess URLs every time they see them. Even if the proposed games might facilitate a raise of awareness and possible behavioral changes, it is not the focus of our current prototypes and might be studied in more detail in the future.

3.2 Learning Goals

To formulate learning goals for our games, we used BRT and structured our goals alongside the six different cognitive processes of the taxonomy: (1) remember, (2) understand, (3) analyze, (4) apply, (5) evaluate and (6) create. Furthermore, BRT provides a distinction of the following knowledge dimensions: (a) factual, (b) conceptual, (c) procedural, and (d) meta-cognitive knowledge. As such, the covered learning content ranges from terminology, concepts, and principles to subject-specific techniques and methods, i.e., URL parsing and manipulation. Hence, our learning goals cover factual, conceptual but also procedural knowledge. For now, meta-cognitive knowledge is not explicitly imparted in the games, i.e., knowledge and awareness of one's own cognition [15].

Table 1: *Learning goals including the mapping to both games (marked with x if a learning goal applies to the analysis game (A) or creation game (C))*

	After playing the learning game, players should be able to ...		A	C
Remember	...	know the structure of URLs by recalling its components.	x	x
	...	name the manipulation techniques for URLs by listing the manipulation techniques for individual components.	x	x
	...	know the manipulation techniques for URLs by describing the manipulation of the components.	x	x
Understand	...	understand the structure of URLs by explaining the purpose of the components.	x	x
	...	understand the manipulation of the structure of URLs by explaining manipulation techniques for the components.	x	x
Apply	...	determine the individual components of a URL by performing URL parsing.	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
	...	detect manipulations in the structure of a URL by identifying manipulated components.	x	
	...	recognize the manipulation technique applied to a URL by identifying/recognizing the manipulated component.	x	
Evaluate	...	assess the correctness of the structure of a URL by checking the components.	x	
	...	assess the manipulation of the structure of a URL by checking the components and identifying manipulated components.	x	
	...	distinguish benign URLs from manipulated URLs by comparing both URLs in terms of applied manipulation(s).	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

4 Implementation

This section describes the implementation of both games, i.e., the tutorial and level design, gameplay, feedback mechanisms, and technical details about the underlying framework and the development.

The first game is called "All sorts of Phish" and is referred to as the *analysis game*[4] since it follows an analytical approach similar to existing games where players have to analyze a given URL. Addressing our first design goal, players have to sort URLs based on the applied manipulation techniques instead of only deciding whether URLs are malicious or benign.

The second game is called "A Phisher's Bag of Tricks" and is referred to as the *creation game*[5]. Here, players are required to apply manipulation techniques to create their own benign and phishing URLs (fulfilling the second design goal).

Overall, the structure of the games follows an alternating tutorial-level scheme where players are first introduced to new knowledge in a tutorial and then continue to practice and acquire desired skills by completing one or more levels. To navigate through the tutorials and levels, a tutor character in the form of an NPC is introduced at the beginning of each game. In the analysis game, the tutor character is embodied by a Roman figure which introduces the URL structure and different manipulation techniques, while the creation game follows the idea of switching roles, which is why a mysterious tech-savvy 'phisher' character is guiding players through tutorials and levels. In both games, digital storytelling is kept to a minimum, and elements and themes could be changed easily by providing different assets (e.g., images for background or characters).

4.1 Tutorial Structure

Both games introduce new concepts in interactive, stepwise tutorials. The respective tutor character guides players through the game by explaining the game mechanics and teaching the required knowledge to progress through the next levels. Players can follow the tutorials in a self-paced manner and can go back to previous steps or tutorials if needed. In the current implementation of the game prototypes, the tutorials do not diverge into digital storytelling beyond the introduction of the tutor characters. This way, tutorials are kept short so players would not be discouraged by lengthy explanations.

Within four tutorials, the analysis game covers the URL structure and key concepts like RD and IP addresses as well as manipulation techniques for RD, subdomains, and paths. In the creation game, in contrast, the learning content is distributed using five tutorials. In the current implementation, the tutorial content of both games is similar but not equal: IP and random URLs are only included in the analysis game since they offer an easy starting point there but do not add meaningful manipulations to the creation game. Similarly, TLDs are an easy, first manipulation technique in the creation game, but are often very complicated to classify without knowledge of the original URL. The differences in the tutorial sections of the games are later discussed in the evaluation of the game prototypes (see Section 7.2).

4.2 Level Structure and Feedback

In both games, tutorial sections are followed by levels where the concepts and knowledge that were introduced in the tutorials are tested. As such, the levels offer an opportunity for practice and assessment where players apply the knowledge gained in the previous tutorial section and measure their performance.

[4] `https://gitlab.com/learntech-rwth/erbse/analysis-game`, accessed on 07.09.2022
[5] `https://gitlab.com/learntech-rwth/erbse/creation-game`, accessed on 07.09.2022

4.2.1 Analysis Game

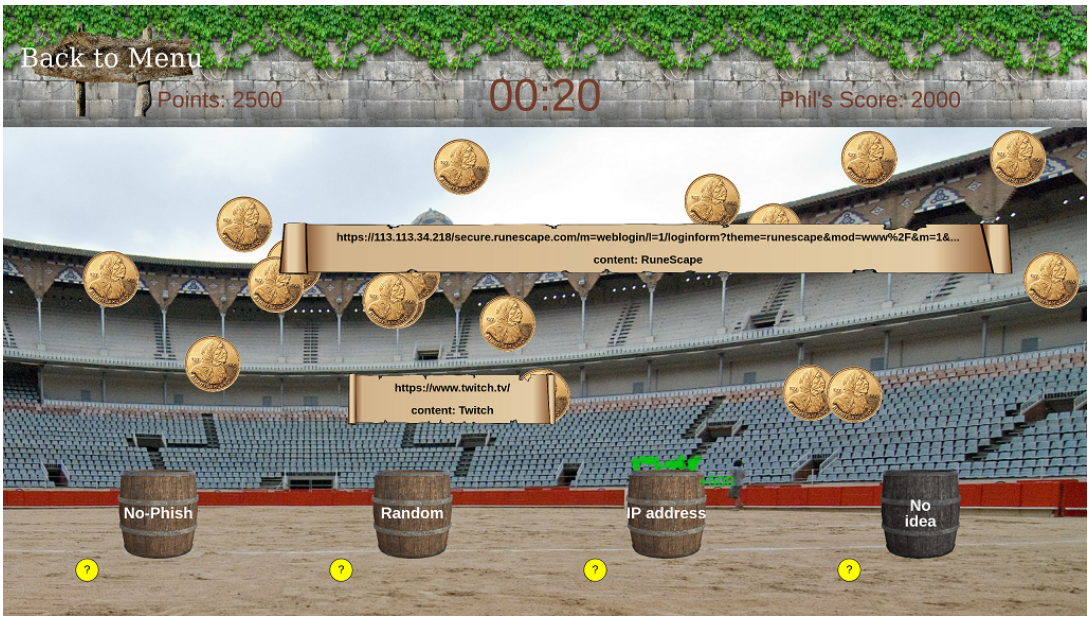

Figure 1: *Level of the Analysis Game. Abstract coins and buckets are used to provide an exemplary theme, however, they can easily be exchanged for other assets.*

A level in the analysis game requires players to analyze given URLs and sort them into different buckets representing different manipulation techniques via drag-and-drop actions (see Figure 1). In the main content area of the level, players are presented with a set of moving coins, which flip upon clicking on them and reveal a URL and a context. The set of URLs is randomized for each player and for each level. For sorting the URLs, players can choose between a bucket for benign URLs and up to five buckets for different phishing URLs: `IP`, `path`, `random`, `RD`, `subdomain`. Lastly, there is an additional bucket (labeled "no-idea") if players are indecisive on how to sort a given URL. This way, players are not forced to make a decision and can discard URLs they do not know. The number of buckets for phishing URLs is increased with each completed tutorial, and they correspond to the different manipulation techniques explained above. The benign and "no-idea" buckets are present from the beginning and remain throughout the game.

Each level lasts a fixed amount of time (default: 90 seconds) and to complete the level successfully, players have to pass a preset score while not making too many mistakes. When dropping a coin into a bucket, players receive instant feedback in the form of a colored aura over the bucket (green: correct; yellow: correct tendency, i.e. a URL classified as phishing but not the correct manipulation technique; red: incorrect; black: discarded) and the increase/decrease of their score (plus points for correct decisions (400 points) or correct tendency (200 points), no points for incorrect decisions or discarded URLs). After the time is up, the game checks whether players achieved a higher score than a preset value (to challenge players to classify as many URLs as possible; default: 2000 points), and that they did not make too many mistakes and achieved a high classification accuracy (to prevent random guessing; default: at least 75% accuracy must be achieved). Players complete a level successfully if both conditions are met. If they do not meet either one of the conditions, they have to repeat the level. To support players in the learning process, feedback is given after the level terminates. Feedback is presented for different types of mistakes (e.g., when URLs were classified as phishing but not as the correct type), as well as selected correct classifications, in

order to support learning from mistakes and try to catch and prevent misconceptions.

4.2.2 Creation Game

In the creation game, levels require players to solve a set of tasks called *presets* by creating their own URLs. For each preset, players are given a task description and a set of URL parts (e.g., “https”, “.”, “com”). The task is displayed at the top of the screen and is focused on the manipulation techniques that were introduced in the previous tutorial. To solve the task, players have to combine URL parts by moving them into the initially empty URL bar in the center of the screen. Within the URL bar, all URL parts can be sorted to complete a valid URL structure. The first level only establishes vocabulary and the basic structure of URLs, while the tasks in the following levels are always based on an original benign service and domain name (e.g., Amazon - amazon.com). Based on this, players are asked to apply manipulation techniques to create their own phishing URLs. For example, players might be asked to create a phishing URL that includes the target name (e.g., “Amazon”) in a subdomain, with a given RD that must not be changed (see Figure 2). To complete the task, players have to combine and drag URL parts into a URL drop area, which shows the current solution URL. In later levels, players are also allowed to create their own URL parts, to provide room for creativity. Creating new URL parts is possible via a keyboard input area, that can be activated to allow players to create arbitrary input.

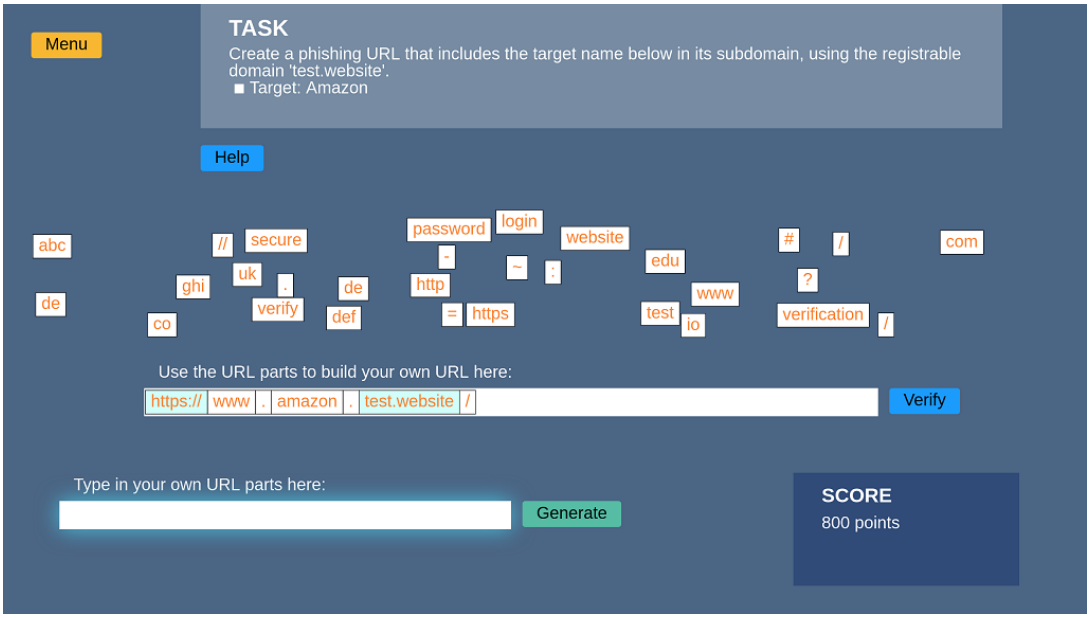

Figure 2: *Level of the Creation Game. The URL bar contains a solution containing predefined (white background) and newly created (turquoise background) URL parts.*

Possible task solutions can be checked by clicking on the corresponding button labeled “Verify”. When a solution is submitted, the game performs a number of automated checks evaluating the created URL with respect to the task of the level. Then, feedback is provided regarding which parts of the task were performed correctly and what has to be changed in case of a failed check. Players get points for passing different types of checks, indicating which tasks are more complicated and opening up the possibility to include non-mandatory challenges, that give additional points but do not have to be completed to advance in the game. As opposed to the analysis game, tasks in the creation game do not have to be completed in a fixed amount of time (default setting), however, depending on the context the game is played

in, a timer can be activated.

4.3 Development

Both games were implemented using the Multi-Touch Learning Game framework[6] (MTLG), a game development framework based on the HTML5-Canvas element and native JavaScript. Originally developed to support game development on multi-touch tabletop displays, MTLG is a modular and versatile game development framework for any browser-capable device supporting any display resolution (e.g. mobile, or desktop). For the evaluation of gameplay and to support game learning analytics, both games implement event logging capabilities to capture the players' actions and game state. Log data is transmitted asynchronously to an external log server for further processing and analysis. For the identification of different players, human-readable, unique player IDs are generated by the log server. This way, we are able to match game sessions to external evaluation instruments (e.g., surveys). Since logging and player IDs are only needed for evaluation purposes in future user studies, they can be deactivated for production environments without evaluation.

4.4 Preliminary Evaluation

Before evaluating the games in a comprehensive user study, a preliminary evaluation was performed serving as functionality testing to find potential issues and bugs. As such, in December 2019, each game was played by a group of participants with the opportunity to give feedback and report problems during gameplay ($N_{\text{Creation}} = 22$, $N_{\text{Analysis}} = 18$). The participant sample consisted of a group of Computer Science (CS) students. Since prior knowledge in CS was expected, the focus of the evaluation was primarily on functionality testing rather than the suitability of the learning content. Feedback from the evaluation was used as input to improve the implementation of the games and resolve presented issues and bugs.

The study was conducted in two one-hour sessions, in which each group played one of the games. Participants played the games and answered a short survey about the functionality and user experience. Participant assignment to either group was done randomly by an external party, without any influence by the instructors of the study. The survey consisted of open-ended questions about positive and negative aspects of the games, bugs and issues that were encountered during gameplay, and possible improvements for further development.

The importance of the topic and how it was motivated, notions of rewards, and the feedback after playing a level were received as positive aspects of the games. Further, participants praised the fun and engaging way to learn about URLs and phishing, the quality of the examples, and detailed explanations. They also emphasized the importance of the topic and that the games could be beneficial to learners without prior knowledge. One participant described the creation game as a fun and engaging experience to learn about phishing, while another participant highlighted how the combination mechanics invited tinkering with the URL manipulations.

Meanwhile, students also identified several bugs, including a game-breaking problem in the creation game that only appeared in a small number of cases due to the random selection of URLs in the task descriptions, and problems on slower devices that made it complicated to open coins to reveal URLs in the analysis game. Participants recommended replacing the coins with buttons or displaying the URLs immediately and without clicking on the coins if the problems could not be resolved. Furthermore, four participants criticized the explanation texts, which were perceived as too long and boring, as well as missing context and the lack of advanced URL manipulation techniques. This was also reflected in the suggested

[6] `https://mtlg-framework.gitlab.io/`, accessed on 07.09.2022

improvements, where some participants proposed interactive elements in our tutorials, which we implemented in updated versions of the games. Other proposed improvements included a competitive multiplayer mode and even more variety of game mechanics.

In addition to reported bugs and problems while playing the games, some participants also noted, that they did not know all of the services that were presented in the analysis game. This indicates potential problems in classifying URLs of services users are not familiar with. We plan to analyze this more closely by taking a look at log data, to find out if unknown services are actually classified with less accuracy, and if this might also have an effect on the overall learning outcome.

4.5 Baseline Implementation for Comparative Evaluation

In order to compare the new game prototypes to existing games with the commonly used binary decision mechanics, a third game prototype was implemented to serve as a baseline implementation in the evaluation. We decided to implement a new baseline game that is similar to existing anti-phishing learning games, as no existing game was able to be adapted and aligned with the design and learning content of the analysis and creation game.

As such, the third learning game prototype is an adapted version of the analysis game and is referred to as the *decision game*[7]. The decision game is a direct clone of the analysis game, with identical tutorial sections and the same alternating tutorial and level structure in which players first learn about a manipulation technique and afterward are challenged in the levels to apply their gained knowledge. While the level design is also very similar, the classification mechanics used in the analysis game is reduced down to a binary decision scheme similar to existing games (e.g., [18, 29, 30]). Therefore, only two buckets for benign and malicious URLs are available. To still allow players to discard URLs they are unable to classify confidently, a third bucket is provided. Although this might extend the decision to be made between three choices, we argue that the decision mechanics are still binary since players are challenged to decide whether a URL is malicious or not. Furthermore, passing a level cannot be achieved by discarding all URLs but rather by deciding correctly and scoring enough points to pass a given score. The remaining implementation is kept similar to the analysis game.

5 User Study

In order to gain insights into the games' effectiveness and to compare the different game mechanics, we designed a user study to evaluate the different game prototypes using the following research questions:

- **RQ-1**: Do the games have a positive influence on the participants' performance and confidence 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**: How do the participants' performance and confidence in classifying URLs change in a longitudinal test?

The overall objective of our research was to evaluate the effects of the new game mechanics and their insights into players' in-game decisions. Thus, we evaluated the games in pre-, post-, and longitudinal tests focused on URL classification knowledge. Furthermore, we compared the games by analyzing the players' performance and confidence after playing either game.

[7] `https://gitlab.com/learntech-rwth/erbse/decision-game`, accessed on 07.09.2022

As such, we designed a user study with three distinct groups of participants, where each group played one game. The study followed a between-group pre-test/post-test design with additional longitudinal testing. While the three games served as independent variables, the performance and confidence in pre-, post-, and longitudinal tests served as our dependent variables.

5.1 Apparatus and Materials

For the different test phases, an online survey containing the following components was used:

- A **URL classification test** that measures the participants' performance and confidence in classifying a set of URLs was utilized in the pre-, post-, and longitudinal tests. For each URL, the participants had to decide whether it was phishing or benign. Further, they had to rate their confidence in the classification decision on a 6-point Likert scale (from "very uncertain" to "very certain"). The URL classification test was utilized in the pre-, post-, and longitudinal tests.
- A **Demographics** questionnaire containing questions regarding age, gender, and educational background was used to be able to describe the participant sample.

The URLs for the URL classification test were generated by collecting benign login URLs of popular websites, which are identical to the set of login pages used in the games (Section 3.1). While the games apply manipulation techniques to create possible but likely non-existent URLs for both benign and phishing examples, actual benign URLs are used in the URL tests of the survey, to avoid confusing participants who exactly remember an existing URL. Similar to the content in the games, different manipulation techniques were applied to the benign login URLs to create various phishing URLs. Overall, 13 phishing URLs, including examples for all manipulation techniques present in the games, and 7 benign URLs were created for the pre-test, with an additional 7 phishing and 3 benign URLs in the post- and longitudinal tests to control for learning bias on the pre-test URLs. In the tests, all participants were shown the same URLs, however, the order of the URLs was randomized to avoid learning bias between the URLs. While the task in the URL classification test is similar to the analysis or decision games, its content is explicitly different from URLs appearing in the games, meaning none of the URLs that appear in the test were included in the games. Further, the analysis game requires players to sort into multiple categories, while the test reduces the task to the binary decision between phishing and benign.

5.2 Procedure

The first part of the study was conducted as a remote lab study using video conferencing software and a web browser on the participants' personal devices. It was structured into five phases with an expected duration of 70 minutes: For the introduction (phase 1), the participants were briefed on the study topic and presented with a definition and example of phishing. Next, participants were given the pre-test survey (phase 2). After completing the pre-test, the survey software directed participants randomly to one of the games (phase 3). After playing, the participants returned to the survey for the post-test (phase 4). Lastly, after the participants completed the survey, a debriefing informed them about the goal of the study and the instructors answered open questions (phase 5).

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 survey. The participants did not know that different games were tested, nor did they know to which group they were assigned. During the session, the participants were asked to start the survey and proceed at their own

pace, as no further instructions were necessary. In case participants had a question or needed technical support, they could contact the instructors and receive help without disrupting other participants.

The second part, the longitudinal study, only contained a survey and did not require additional supervision. As such, participants were invited to participate in the longitudinal study three months after the original study, and to fill out the longitudinal survey independently.

5.3 Participants

The study was conducted in two parts, with 88 participants in November 2020 who played the analysis ($N_A = 40$) and creation ($N_C = 48$) games, and 45 participants in May 2021 who played the decision game. For recruiting, the study was advertised 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. Advertisement for the study was done online by posting about the study in different social network groups of universities as well as distributing it via university mailing lists. A financial incentive of 15 EUR was offered to each participant since the study required active participation for 60-70 minutes. For participation in the longitudinal test three months later, a lottery of 4×10 EUR was offered to all participants who completed the additional survey. Both the methods for recruiting and the financial incentives may have introduced a potential selection bias (see Section 7.1).

Among the participants, 36.09% identified as male and 63.91% as female. The majority were students, with 81.95% of the participants reporting their highest degree to be either a Bachelor's degree or high school diploma. The remaining participants had mainly either completed a Master's degree (12.03%) or vocational training (3.01%). The majority of participants were between 20-29 years old (78.20%).

For the longitudinal test three months after the pre-test/post-test part of the study, we experienced a dropout of 52.63%, leading to a response rate of only 63 participants ($N_A = 17$, $N_C = 25$, $N_D = 21$). This limits the evaluation of longitudinal effects and calls for reproduction with a larger sample.

6 Results

In the following, we present the results of a series of analyses and tests conducted to answer the research questions described in Section 5. As such, we consider three groups depending on which game the participants played: the *creation game group*, the *analysis game group*, and the *decision game group*. In general, a significance level of $\alpha = .05$ was used. Based on the pre-, post-, and longitudinal testing, we computed performance scores as relative scores, i.e. number of correctly classified URLs divided by the number of all URLs. The confidence levels are computed by the average of confidence ratings (ranging from 1 to 6) in the different instances of the URL classification test. Both performance scores and confidence levels are measured using an interval scale.

6.1 Differences between Pre- and Post-Tests

In response to **RQ-1**, we focus on the general effectiveness of the games by comparing the participants' performance scores and confidence levels in pre- and post-test. As shown in Table 2, both the mean performance score and the mean confidence level improved between pre- and post-test for all games, meaning that the participants' performance and confidence in classifying URLs increased after playing either one of the games. Since the performance and confidence means of newly added URLs in the longitudinal test (*post-new*) are similar to

the means of URLs already tested in the pre-test, we argue that a potential learning bias on reoccurring URLs is negligible.

Table 2: *Means (and standard deviation) for performance and confidence in pre- and post-test including means on partial URL sets. The indices "pre" and "post" are used to distinguish pre- and post-tests, and hyphenated suffixes in the indices are used to distinguish test scores on the URLs also used in the pre-test (post-pre) or newly added URLs (post-new).*

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

Next, we performed a one-tailed Student's t-test for each game, comparing the results of the classification task on pre-test URLs between pre- and post-test. As shown in Table 3, the results indicate, that the participants' performance and confidence increased significantly in the post-test for all three games. There are, however, differences in effect sizes, that are large for performance scores for the analysis game and the decision game, but not for the creation game. In the tests for confidence, effect sizes are large for all games.

Table 3: *Results of t-tests comparing performance and confidence in pre- and post-test for all three games. If a deviation from normality was detected, a Wilcoxon signed-rank test was used (Shapiro-Wilk test, cut-off value $\alpha < 0.05$).*

	performance			confidence		
Game	Test statistic	p-value	effect size	Test statistic	p-value	effect size
Analysis	$t(39) = -6.404$	$p < .001$	$d = -1.013$	$W = 1.000$*	$p < .001$	$r = -0.997$
Creation	$t(47) = -3.459$	$p < .001$	$d = -0.499$	$t(47) = -7.850$	$p < .001$	$d = -1.133$
Decision	$W = 24.500$*	$p < .001$	$r = -0.946$	$t(44) = -10.273$	$p < .001$	$d = -1.531$

*Deviation from normality detected.

6.2 Differences between Games

For **RQ-2**, we evaluated the differences between the three game prototypes by comparing the participants' performance and confidence in the post-test. As shown in Table 2, mean values seem to suggest, that players of the analysis and decision games performed similarly well, while participants of the creation game performed worse.

We compared the performance scores in the post-test for all three games using an ANCOVA with the games as the between-group factor, the post-test performance as the dependent variable, and the pre-test performance as a covariate. After checking for homogeneity (Levene, $F(2, 130) = 1.207, p = 0.302$), the ANCOVA did 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 ($F(1, 129) = 45.333, p < 0.001, \eta_p^2 = .260$).

Similarly, we compared the participants' confidence levels for all three games by using an ANCOVA as well. Here, the results reveal significant differences between the games in the post-test ($F(2, 129) = 5.429, p = .005, \eta_p^2 = .078$) and for the pre-test as covariate ($F(1, 129) = 79.372, p < 0.001, \eta_p^2 = .381$). Post-hoc tests using Holm-correction show significant differences between the creation game and both the analysis game and the decision game (both $p_{\text{Holm}} = .015$).

6.3 *Differences in Longitudinal Testing*

For the evaluation of long-term effects, we can only consider participants who completed pre-, post-, and longitudinal tests, which results in a reduced participant sample due to a dropout of 52.63%.

As shown in Table 4, both performance and confidence means indicate a decline between the post- (*post-pre*) and longitudinal test (*long-pre*), with the pre-test performance score remaining the lowest for all games. Similar to the post-test means, we argue that a potential learning bias on reoccurring URLs is negligible since the performance and confidence means of newly added URLs in the longitudinal test (*long-new*) are similar to the means of URLs already tested in the pre- and post-test.

Table 4: *Performance means in longitudinal test (with index "long") and comparison to pre- and post-test scores.*

		performance				confidence			
Game	N	$M_{\text{pre}}\ (SD)$	$M_{\text{post-pre}}\ (SD)$	$M_{\text{long-pre}}\ (SD)$	$M_{\text{long-new}}\ (SD)$	$M_{\text{pre}}\ (SD)$	$M_{\text{post-pre}}\ (SD)$	$M_{\text{long-pre}}\ (SD)$	$M_{\text{long-new}}\ (SD)$
Analysis	17	.679 (.095)	.847 (.087)	.812 (.070)	.782 (.119)	4.088 (.705)	4.968 (.497)	4.676 (.405)	4.629(.535)
Creation	25	.698 (.130)	.762 (.105)	.754 (.131)	.740 (.119)	4.166 (.702)	4.720 (.649)	4.578 (.709)	4.764 (.703)
Decision	21	.686 (.107)	.821 (.092)	.774 (.119)	.771 (.096)	4.186 (.821)	4.993 (.478)	4.700 (.739)	4.676 (.728)

To evaluate the performance differences between pre-, post-, and longitudinal testing, a repeated-measures ANOVA was performed, using the three tests as repeated measures and the games as the between-subject factor. The results of the ANOVA confirm that there are significant differences of participants' performances between the pre-, post-, and longitudinal tests ($F(1.810, 120) = 44.727$, $p < .001$, $\eta_p^2 = .427$). Post-hoc tests with Holm correction show that pre-test performance is significantly lower than both post- and longitudinal test performances for the analysis game and the decision game ($p < .020$ for all tests), but not for the creation game ($p = .357$ for the comparison of pre- and longitudinal performance scores). Further, performance differences between the post- and longitudinal tests are nonsignificant for all games.

Similar results can be observed for the participants' confidence levels in pre-, post-, and longitudinal tests. Here, another repeated-measures ANOVA was performed, using the three tests as repeated measures and the games as between-subject factors. Since Mauchly's test for sphericity was violated ($\chi^2(2) = 19.082$, $p < .001$), the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\epsilon = .783$). The results confirm that there are significant differences in the participants' confidence between the pre-, post- and longitudinal tests ($F(1.567, 90.019) = 35.854$, $p < .001$, $\eta_p^2 = .374$). Furthermore, post-hoc tests with Holm correction show that confidence in the pre-test is significantly lower than in post- and longitudinal tests for the analysis game and the decision game ($p < .023$ for all tests), but not for the creation game ($p = .357$ for the comparison of confidences in pre- and longitudinal test). The confidence differences between post- and longitudinal tests are again nonsignificant for all games.

7 Discussion

As presented in the previous section, the results show that all three games led to a significant increase in the participants' performance and confidence (RQ-1) and that the performance increase was still visible in the longitudinal test results (RQ-3). Overall, the chosen game design and the decision to teach URL parsing in a structured approach together with manipulation techniques proved to be effective. On the other hand, the comparison of the three games results in significant differences in confidence levels (RQ-2), while the participants'

mean performance scores only differ slightly between the games. In the following, we discuss the study setup and how it could have influenced the outcomes, followed by a discussion of the study results and their implications.

7.1 Study Setup

The participants of our user study were recruited using channels that are mainly frequented by students in higher education. As a result, the majority of participants were students, which leads to concerns regarding the generalizability of results. In particular, we argue that, while results may be generalized to the population of students, the study should be repeated with a sample more closely representing the general population for further generalization.

Additionally, we experienced a large dropout of participants for the longitudinal tests. This not only makes statistical tests less powerful but might also introduce a self-selection bias, as only participants who were interested in the topic decided to participate in the longitudinal test. The generally higher means of post-test scores in the longitudinal test (see Table 4) compared to the means of all post-test participants (see Table 2) seem to support the hypothesis that self-selection took place to a certain amount. This might have led to further deviations of the participants from the general population, and might be counteracted in the future by changing the study design.

The study was performed as an online lab study, with a focus on knowledge about phishing URLs. As such, the study results, in particular the positive effect on classification accuracy, only indicate changes in our test setup. We did not evaluate situational phishing awareness, nor the participants' ability to detect phishing websites in a realistic scenario, e.g., in a simulated phishing attack. Future work might integrate information on situational awareness into the games, and evaluate whether the knowledge learned in the games can be applied in the real world by including simulated attacks after playing the games. Further testing for behavioral change would require detailed user supervision or assessment, to understand how users evaluate not only potential phishing URLs but also benign URLs. However, the learning goals of our games only cover knowledge and skills on reading and assessing URLs, not the strategies on when to apply them. In order to foster a behavioral change, we propose either extending the game to include such strategies or encouraging users to develop their own strategies that fit their daily habits. Still, we argue that even the conveyed knowledge on URLs can be applied in a real-world setting, e.g., when users are unsure about an encountered URL and need to assess its origin. Here, users can actively decide to avoid clicking on URLs they consider suspicious.

7.2 Study Results

For RQ-1, we found that both performance scores and confidence levels increased significantly for all games, with high effect sizes in both games, which validates the suitability of our games in our study setup. In particular, problems that were encountered in the preliminary study were no longer present.

Regarding RQ-2, we found that differences in mean performance scores of the three games were not significant. This could either indicate, that the changes in game mechanics did not have an effect on the learning outcomes, only on confidence, or that there are changes that our study setup was unable to capture. While the games were designed similarly and with a final comparison in mind, the creation game differs partially from the other two games in regards to the covered content and its presentation in the tutorials, which complicates the comparison. While we restricted the test in RQ-2 to URL categories that appear in both games, these differences might still have influenced the outcomes. Newer, improved versions of the games,

with more closely aligned learning content, have already been created and might be compared in the near future to eliminate the effect of differences in the tutorial sections.

The participants' confidence levels, on the other hand, did result in significant differences between the games, as players of the creation game were on average less confident than players of the other games in the post-test. This could be due to the game mechanics in the creation game being perceived as more complex, or the players' self-reported confidence levels might be more accurate for the creation game (i.e. players of the analysis and decision games overestimated their classification accuracy). We did not pursue this question further in the current paper but would recommend evaluating the difficulty of the games as well as the suitability of our instruments to capture confidence levels.

The direct comparison of the analysis and decision games, which only differ in the complexity of the classification task, did not result in significant differences in performance scores or confidence levels. Here, it is possible that the tutorials structured around URL parts already sufficiently prepared players for the classification task, and that the more elaborate practice provided by the game mechanics in the analysis game does not have a susceptible effect on learning outcomes. Future studies might test whether the game mechanics of the analysis game have an effect on URL parsing abilities beyond the binary decision required in the URL tests, e.g. by having participants select the registrable domain, path, or subdomains of a given URL. While the study did not signify immediate advantages in the URL classification task, we still argue that the in-game log data which is generated when using the more complex game mechanics offers more detailed insights into players' behaviors, and might be used to better detect and prevent misconceptions in the future. These results might be transferable beyond the domain of anti-phishing learning games. In particular, it is likely that the additional insights into players' decision processes and the structured tutorial approach may benefit learning games that currently rely on binary decisions in other contexts as well. Future work could explore a game analysis using the LM-GM model [21] to further align learning and game mechanics. It may also assess, how perceived enjoyment changes when varying game mechanics, or whether additional game mechanics could increase the mapping of real-world tasks and learning activities with gameplay.

As for RQ-3, results indicate that knowledge can be retained over a three-month period since participants' performances and confidences did not decline significantly and were still higher than in the pre-test. Future work might evaluate knowledge retention over even longer time periods as well as the need for periodical replay of the games. This may necessitate a redesign of the games to allow for returning players as they might not need to reiterate over all tutorials and levels.

In all, we found that playing any of the three games resulted in significant improvements, which were at least partly retained over a three-month period. While there are some indications that game mechanics might have an effect on the classification accuracy in the URL tests, we did not find this difference to be significant in our study setup. Still, enhanced log data that was produced by the more complex game mechanics led to more insights on the players' in-game behavior, which might be utilized to better understand and prevent misconceptions, give feedback during the games, or even change the learning content to better fit the player. While our work serves as a case study of understanding how redesigning game mechanics in anti-phishing learning games may improve players' performance in identifying phishing URLs, it may also serve the more general research on alignment of learning and game mechanics, similar to [22–24].

8 Conclusion and Future Work

In this paper, we presented the design and evaluation of two new anti-phishing learning games. While existing games are characterized by a lack of game mechanics covering more than factual and conceptual knowledge as well as higher-order cognitive processes, the presented games follow both a more elaborate analytical approach and a constructive approach utilizing alternative game mechanics. A case study with 133 participants confirms the games' effectiveness in improving the participants' URL classification performance and self-reported confidence. While the proposed game mechanics did not lead to immediate improvements compared to a baseline implementation, we argue that there are still advantages to applying the proposed game mechanics. Further, a longitudinal study ($N = 63$) confirmed that knowledge was at least partially retained three months after playing the games. Our results indicate, that the learning content and game design of the games is effective in teaching knowledge about the structure of URLs and different manipulation techniques. They pose the question of whether more complex game mechanics might be utilized to improve feedback and adaptation in the future. While this work focuses on the domain of anti-phishing education, similar issues with limitations in game design and assessment capabilities might be observable in other domains as well. As such, future work on linking game mechanics with learning and assessment more accurately may provide more detailed insights into players' learning processes [31].

Acknowledgements

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

References

[1] E. E. Lastdrager, "Achieving a consensual definition of phishing based on a systematic review of the literature," *Crime Science*, vol. 3, no. 1, p. 9, Sep. 2014. doi: 10.1186/s40163-014-0009-y

[2] 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

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

[4] A. Chattopadhyay, C. Maschinot, and L. Nestor, "Mirror Mirror On The Wall - What Are Cybersecurity Educational Games Offering Overall: A Research Study and Gap Analysis," in *IEEE Frontiers in Education Conference*, ser. FIE '21. IEEE, Oct. 2021. doi: 10.1109/FIE49875.2021.9637224. ISSN 2377-634X pp. 1–8.

[5] K. Köhler, R. Röpke, and M. R. Wolf, "Through a Mirror Darkly – On the Obscurity of Teaching Goals in Game-Based Learning in IT Security," in *Simulation & Gaming Through Times and Across Disciplines*. Warsaw: Akademia Leona Kozminskiego, 2019, pp. 324–335.

[6] 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, 2020. doi: 10.1007/978-3-030-62433-0_3. ISBN 978-3-030-62433-0 pp. 41–60.

[7] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, "A systematic literature review of empirical evidence on computer games and serious games," *Computers & Education*, vol. 59, no. 2, pp. 661–686, 2012. doi: 10.1016/j.compedu.2012.03.004

[8] E. A. Boyle, T. Hainey, T. M. Connolly, G. Gray, J. Earp, M. Ott, T. Lim, M. Ninaus, C. Ribeiro, and J. Pereira, "An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games," *Computers & Education*, vol. 94, pp. 178–192, Mar. 2016. doi: 10.1016/j.compedu.2015.11.003

[9] F. Alotaibi, S. Furnell, I. Stengel, and M. Papadaki, "A Review of Using Gaming Technology for Cyber-Security Awareness," *Information Security Research*, vol. 6, no. 2, pp. 660–666, 2016. doi: 10.20533/ijisr.2042.4639.2016.0076

[10] A. L. Compte, D. Elizondo, and T. Watson, "A renewed approach to serious games for cyber security," in *International Conference on Cyber Conflict: Architectures in Cyberspace*. Tallinn: IEEE, 2015. doi: 10.1109/CYCON.2015.7158478 pp. 203–216.

[11] M. Hendrix, A. Al-Sherbaz, and V. Bloom, "Game Based Cyber Security Training: Are Serious Games suitable for cyber security training?" *Serious Games*, vol. 3, no. 1, pp. 53–61, 2016. doi: 10.17083/ijsg.v3i1.107

[12] R. Roepke and U. Schroeder, "The Problem with Teaching Defence against the Dark Arts: A Review of Game-based Learning Applications and Serious Games for Cyber Security Education," in *International Conference on Computer Supported Education*, vol. 2. Heraklion: SciTePress, 2019. doi: 10.5220/0007706100580066 pp. 58–66.

[13] V. Pastor, G. Díaz, and M. Castro, "State-of-the-art simulation systems for information security education, training and awareness," in *IEEE EDUCON Conference*. Madrid: IEEE, 2010. doi: 10.1109/EDUCON.2010.5492435 pp. 1907–1916.

[14] J.-N. Tioh, M. Mina, and D. W. Jacobson, "Cyber security training a survey of serious games in cyber security," in *IEEE Frontiers in Education Conference (FIE)*. Indianapolis: IEEE, 2017. doi: 10.1109/FIE.2017.8190712. ISBN 978-1-5090-5920-1 pp. 1–5.

[15] D. R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview," *Theory Into Practice*, vol. 41, no. 4, pp. 212–218, Nov. 2002. doi: 10.1207/s15430421tip4104_2

[16] 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 *Symp. on Usable Privacy and Security*, ser. SOUPS âĂŹ07. New York: ACM, 2007, pp. 88–99.

[17] P. Weanquoi, J. Johnson, and J. Zhang, "Using a game to improve phishing awareness," *Cybersecurity Education, Research and Practice*, vol. 2018, no. 2, p. 2, 2018.

[18] G. Canova, M. Volkamer, C. Bergmann, and R. Borza, "NoPhish: An Anti-Phishing Education App," in *Security and Trust Management (STM)*, vol. 8743. Cham: Springer, 2014. doi: 10.1007/978-3-319-11851-2_14. ISBN 978-3-319-11850-5 pp. 188–192.

[19] M. Romero, L. Dumont, S. Barma, S. Daniel, M. Ferrer, V. Hénaire, M.-B. Răileanu, A. Lepage, B. Lille, and A. Patino, *Digital Games and Learning*. JFD, 2020. ISBN 978-2-89799-055-8

[20] G. Kalmpourtzis and M. Romero, "Constructive alignment of learning mechanics and game mechanics in Serious Game design in Higher Education," *International Journal of Serious Games*, vol. 7, no. 4, pp. 75–88, Dec. 2020. doi: 10.17083/ijsg.v7i4.361

[21] S. Arnab, T. Lim, M. B. Carvalho, F. Bellotti, S. de Freitas, S. Louchart, N. Suttie, R. Berta, and A. De Gloria, "Mapping learning and game mechanics for serious games analysis," *British Journal of Educational Technology*, vol. 46, no. 2, pp. 391–411, 2015. doi: 10.1111/bjet.12113

[22] M. Callaghan, M. Savin-Baden, N. McShane, and A. G. Eguíluz, "Mapping Learning and Game Mechanics for Serious Games Analysis in Engineering Education," *IEEE*

Transactions on Emerging Topics in Computing, vol. 5, no. 1, pp. 77–83, Jan. 2017. doi: 10.1109/TETC.2015.2504241

[23] M. J. Callaghan, N. McShane, A. G. Eguíluz, T. Teillès, and P. Raspail, "Practical application of the Learning Mechanics-Game Mechanics (LM-GM) framework for Serious Games analysis in engineering education," in *2016 13th International Conference on Remote Engineering and Virtual Instrumentation (REV)*, Feb. 2016. doi: 10.1109/REV.2016.7444510 pp. 391–395.

[24] A. Patino, M. Romero, and J.-N. Proulx, "Analysis of Game and Learning Mechanics According to the Learning Theories," in *2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, Sep. 2016. doi: 10.1109/VS-GAMES.2016.7590337 pp. 1–4.

[25] J. Plass, B. Homer, C. Kinzer, J. Frye, and K. Perlin, "Learning Mechanics and Assessment Mechanics for Games for Learning," Games for Learning Institute, New York, Tech. Rep., 2011.

[26] 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)*, 2018. doi: 10.1109/ECRIME.2018.8376206. ISSN 2159-1245 pp. 1–12.

[27] 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 *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019. doi: 10.1145/3319535.3363188 pp. 2489–2504.

[28] 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 *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. doi: 10.1145/3313831.3376298 pp. 1–12.

[29] 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 *Proceedings of the 3rd Symposium on Usable Privacy and Security*, ser. SOUPS '07. Association for Computing Machinery, 2007. doi: 10.1145/1280680.1280692. ISBN 978-1-59593-801-5 pp. 88–99.

[30] N. A. G. Arachchilage, S. Love, and C. Maple, "Can a mobile game teach computer users to thwart phishing attacks?" *Journal of Infonomics*, vol. 6, no. 3/4, pp. 720–730, 2015. doi: 10.20533/iji.1742.4712.2013.0083

[31] P. Lameras, S. Arnab, I. Dunwell, C. Stewart, S. Clarke, and P. Petridis, "Essential features of serious games design in higher education: Linking learning attributes to game mechanics," *British Journal of Educational Technology*, vol. 48, no. 4, pp. 972–994, 2017. doi: 10.1111/bjet.12467