Extending Game-Based Anti-Phishing Education using Personalization

Design and Implementation of a Framework for Personalized Learning Game Content in Anti-Phishing Learning Games

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

René Christian Röpke, M.Sc.

aus Frankfurt am Main

Berichter: Univ.-Prof. Dr.-Ing. Ulrik Schroeder
Univ.-Prof. Dr.-Ing. habil. Ulrike Lucke

Tag der mündlichen Prüfung: 12. April 2023

RWTH Aachen University

Faculty of Mathematics, Computer Science and Natural Sciences
Learning Technologies Research Group

Dissertation

Extending Game-Based Anti-Phishing Education using Personalization

Design and Implementation of a Framework for Personalized Learning Game Content in Anti-Phishing Learning Games

Author:
René Christian Röpke, M.Sc.

Reviewer:
Univ.-Prof. Dr.-Ing. Ulrik Schroeder
Univ.-Prof. Dr.-Ing. habil. Ulrike Lucke

Dissertation submitted to the Faculty of Mathematics, Computer Science and Natural Sciences at RWTH Aachen University to obtain the academic degree of Doctor of Natural Sciences

DOI: 10.18154/RWTH-2023-04991

Abstract

Phishing poses an imminent and wide-ranging threat to Internet users worldwide, in which attackers use methods of deception to lure victims into disclosing information. Recent reports state high numbers of phishing incidents and, so far, technical solutions fail to stop the threat completely. As a complementary approach, user education using anti-phishing learning games has been explored to raise awareness and teach the necessary knowledge and skills to detect and protect against phishing attacks. A common game mechanic used in existing games requires learners to classify URLs as either legitimate or phishing in a binary decision scheme. Here, a problem can occur if learners do not know the service of a given URL and are unable to classify the URL due to a lack of reference. As such, learners may revert to guessing which may weaken the game's potential for practice, since learners cannot relate between correct classifications and the applied knowledge. Furthermore, the possibilities for feedback are limited since the binary decision mechanic does not provide any insights into learners' decision processes and possible misconceptions. In this dissertation, the limitations for feedback as well as the problem with classifying unknown URLs in anti-phishing learning games are addressed as follows: First, a review of existing learning games provides insights into their design and covered learning content. Its results are used in guiding the design and implementation of two new game prototypes. Here, the first game extends the before-mentioned binary decision mechanic and requires learners to sort URLs into one of many categories, depending on which manipulation technique was applied to a distinct part of the URL. The second game requires learners to apply different manipulation techniques and create their own malicious URLs using a puzzle mechanic. Next, the means of personalization for anti-phishing learning games are explored and a personalization pipeline is developed. By considering the learners' familiarity with different services and dynamically creating benign and phishing URLs, the content of anti-phishing learning games can be personalized. To evaluate the new game prototypes as well as the application of the personalization pipeline, two comparative user studies are conducted in a between-group design with pre-, post- and longitudinal testing. In the first user study with 133 participants, both games are evaluated and compared to a baseline implementation. While participants of the new games did not perform significantly better than the control group, results show significant improvements in the participants' performance and confidence between pre- and post-tests for all games, as well as notable differences when classifying URLs of unknown and known services. In the second user study with 49 participants, the personalization pipeline is integrated into one of the games, in order to compare its personalized and non-personalized version. Here, personalization enables the control of service familiarity and allows insights into how URLs of unknown services are handled within the game. While participants of the personalized game did not outperform the participants of its non-personalized version, the evaluation of in-game behavior provides insights into learners' decision processes and possible problems or misconceptions. Furthermore, results of a longitudinal evaluation of all games and versions show that knowledge is retained since the participants perform still significantly better than in the pre-test. In all, this dissertation presents first approaches and research results in the domain of personalized anti-phishing learning games. Future work may entail redesigning anti-phishing learning games to incorporate further means of personalization and to understand how learner characteristics can be utilized in anti-phishing learning games.

Zusammenfassung

Phishing stellt eine unmittelbare und weitreichende Bedrohung für Internetnutzer auf der ganzen Welt dar, bei der Angreifer durch Täuschung ihre Opfer dazu verleiten, Informationen preiszugeben. Aktuelle Berichte berichten über eine Vielzahl von Phishing-Angriffen, und bislang konnten die technischen Lösungen die Bedrohung nicht vollständig aufhalten. Als komplementärer Ansatz wird die Schulung der Nutzerinnen und Nutzer mit Hilfe von Anti-Phishing-Lernspielen untersucht, um diese zu sensibilisieren und ihnen die notwendigen Kenntnisse und Fähigkeiten zu vermitteln, Phishing-Angriffe zu erkennen und sich davor zu schützen. Eine gängige Spielmechanik in existierenden Spielen fordert von Lernenden, dass sie URLs in einem binären Entscheidungsschema entweder als legitim oder als Phishing klassifizieren. Hierbei kann ein Problem auftreten, wenn die Lernenden den Dienst einer gegebenen URL nicht kennen und mangels Referenz nicht in der Lage sind, die URL zu klassifizieren. So können die Lernenden auf Raten ausweichen, was das Übungspotenzial des Spiels schmälert, da Lernende keine Verknüpfung zwischen richtigen Klassifizierungen und dem angewandten Wissen schaffen können. Darüber hinaus sind die Möglichkeiten für Feedback begrenzt, da die binäre Entscheidungsmechanik keinen Einblick in die Entscheidungsprozesse und möglichen Fehlvorstellungen der Lernenden gibt. In dieser Dissertation werden die Limitationen für Feedback sowie das Problem der Klassifizierung unbekannter URLs in Anti-Phishing-Lernspielen wie folgt adressiert: Zunächst wird ein Überblick über existierende Lernspiele erarbeitet, der Einblicke in deren Design und die behandelten Lerninhalte gibt. Die Erkenntnisse dienen der Gestaltung und Implementierung von zwei neuen Spielprototypen. Hier erweitert das erste Spiel die zuvor erwähnte binäre Entscheidungsmechanik und verlangt von den Lernenden, URLs in eine von mehreren Kategorien einzuordnen, je nachdem, welche Manipulationstechnik auf einen bestimmten Teil der URL angewendet wurde. Beim zweiten Spiel müssen Lernende verschiedene Manipulationstechniken anwenden und mithilfe einer Puzzlemechanik ihre eigenen bösartigen URLs erstellen. Als nächstes werden die Möglichkeiten zur Personalisierung von Anti-Phishing-Lernspielen betrachtet und eine Personalisierung-Pipeline entwickelt. Durch die Berücksichtigung der Bekanntheit der Lernenden mit unterschiedlichen Diensten und der dynamischen Erstellung von gutartigen und Phishing URLs kann der Inhalt von Anti-Phishing-Lernspielen personalisiert werden. Um die neuen Spielprototypen sowie den Einsatz der Personalisierungs-Pipeline zu evaluieren, werden zwei vergleichende Nutzerstudien in einem Between-Group-Design mit Prä-, Post- und Langzeittests durchgeführt. In der ersten Nutzerstudie mit 133 Teilnehmern werden beide Spiele bewertet und mit einer Referenzimplementierung verglichen. Während die Testpersonen der neuen Spiele nicht signifikant besser abschnitten als die Kontrollgruppe, zeigen die Ergebnisse für alle Spiele signifikante Verbesserungen in der Leistung und im Selbstvertrauen der Testpersonen zwischen Vor- und Nachtest sowie auffällige Unterschiede bei der Klassifizierung von URLs unbekannter und bekannter Dienste. In der zweiten Nutzerstudie mit 49 Testpersonen wird die Personalisierungs-Pipeline in eines der Spiele integriert, um dessen personalisierte und nicht-personalisierte Version zu vergleichen. In diesem Fall ermöglicht die Personalisierung die Steuerung der Bekanntheit der Dienste und gibt Aufschluss darüber, wie URLs von unbekannten Diensten im Spiel gehandhabt werden. Obwohl die Testpersonen des personalisierten Spiels nicht besser abschnitten als die Teilnehmer der nicht-personalisierten Version, bietet die Auswertung des Spielverhaltens Einblicke in die Entscheidungsprozesse der Lernenden und mögliche Probleme oder Fehlvorstellungen. Darüber hinaus zeigen die Ergebnisse einer Langzeitevaluation aller Spielprototypen und -versionen, dass das Wissen erhalten bleibt, da die Testpersonen immer noch deutlich besser abschneiden als im Prätest. Insgesamt stellt diese Dissertation erste Ansätze und Forschungsergebnisse im Bereich der personalisierten Anti-Phishing-Lernspiele vor. Zukünftige Arbeiten könnten die Neugestaltung von Anti-Phishing-Lernspielen umfassen, um weitere Möglichkeiten der Personalisierung einzubeziehen und um zu verstehen, wie die Lernereigenschaften in Anti-Phishing Lernspielen berücksichtigt werden können.

Acknowledgements

In crafting this dissertation, I have been fortunate to receive the guidance, support, and encouragement of many individuals. Without their collective wisdom, dedication, and inspiration, this work would not have been possible. It is with immense gratitude that I acknowledge their contributions to this personal and professional project.

First and foremost, I would like to extend my deepest appreciation to Prof. Ulrik Schroeder, my first supervisor, whose expertise, patience, and unwavering support has been invaluable throughout this process. His guidance and mentorship have not only enriched my research but also significantly fostered my growth as a scholar. My sincerest thanks go also to my second supervisor, Prof. Ulrike Lucke, who supported me and my endeavors with constructive feedback and guiding questions.

Next, my heartfelt thanks go to my colleague and tandem partner Vincent Drury, whose camaraderie has been an essential component of this project's success. The stimulating discussions and exchange of ideas have greatly contributed to the development and depth of my research. This collaboration greatly impacted my personal and professional growth, and is something I wish every doctoral student would have.

Furthermore, I would like to thank Prof. Ulrike Meyer for allowing me to join the ERBSE research project and her team. I am deeply grateful for her guidance and support over the years. As part of this project, I am also thankful to the team of student assistants supporting the game development as well as collaborating in different publications. Their manpower was substantial to my research and the results we achieved.

To my colleagues and friends at the Learning Technologies Research Group, I was lucky to have spent all these countless hours on discussing research, fuzzing about reviewer comments, traveling to conferences, theorizing about oddly tasting Mensa lunches, ranting about teaching, eating too much cake or simply spending an amazing time in the (home) office. They made this team a family and their support in this endeavor was simply incomparable.

I am also deeply grateful for my friends in Aachen and back home in Oberursel and Darmstadt. They were there when this project took all my energy and left me stranded in desperate need of leisure time and comfort. They enriched my life by providing an escape full of fun, laughter and all the silliness one like me needs to recharge. I am thankful for their belief in me and their willingness to listen and share my triumphs and challenges.

Lastly, I owe a debt of gratitude to my family, especially my parents, whose love, support, and belief in me have been my strongest pillars. Their sacrifices, encouragement, and unwavering faith in my abilities have sustained me throughout this challenging journey. I am deeply grateful for all their guidance in life and the opportunities they have provided.

In dedicating this dissertation to all of these remarkable individuals, I acknowledge the invaluable roles they have played in shaping my academic and personal growth. I am truly honored to have had the privilege of learning from and working alongside each of them.

Fakultät für
Mathematik, Informatik
und Naturwissenschaften

RWTH AACHEN UNIVERSITY

Eidesstattliche Erklärung
Declaration of Authorship

I, René Christian Röpke,

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Hiermit erkläre ich an Eides statt / I do solemnely swear that:

1. This work was done wholly or mainly while in candidature for the doctoral degree at this faculty and university;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others or myself, this is always clearly attributed;
4. Where I have quoted from the work of others or myself, the source is always given. This thesis is entirely my own work, with the exception of such quotations;
5. I have acknowledged all major sources of assistance;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published before as listed below.

Aachen, 05.05.2023

- René Christian Röpke -

Pre-Released Publications

This dissertation contains parts of the following pre-released publications. The content and structure was adapted to fit the overall scope and purpose of this dissertation. For my contribution to each of these publications, see the Statement of Originality.

[A] D. Bayrak, R. Röpke, and U. Schroeder. "Konzeption und Entwicklung eines interaktiven E-Mail-Interface für Anti-Phishing Lernspiele". In: *DELFI 2021*. Die 19. Fachtagung Bildungstechnologien. Ed. by A. Kienle, A. Harrer, J. M. Haake, and A. Lingnau. DELFI '21. Bonn: Gesellschaft für Informatik e.V., 2021, pp. 157–162. DOI: 10.18154/RWTH-2021-00081.

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

[C] K. Köhler, R. Röpke, and M. Wolf. "Through a Mirror Darkly – On the Obscurity of Teaching Goals in Game-Based Learning in IT Security". In: *Simulation Gaming Through Times and Disciplines*. International Simulation and Gaming Association Conference. Ed. by M. Wardaszko, S. Meijer, H. Lukosch, H. Kanegae, W. C. Kriz, and M. Grzybowska-Brzezińska. ISAGA '19. Cham: Springer International Publishing, 2021, pp. 61–73. DOI: 10.1007/978-3-030-72132-9_6.

[D] R. Röpke and V. Drury. *Evaluation Data of Anti-Phishing Learning Games*. 2023. DOI: 10.17605/OSF.IO/3YDW9.

[E] R. Roepke, K. Koehler, V. Drury, U. Schroeder, M. 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*. Ed. by G. Hatzivasilis and S. Ioannidis. MSTEC '20. Cham: Springer International Publishing, 2020, pp. 41–60. DOI: 10.1007/978-3-030-62433-0_3.

[F] R. Roepke, U. Schroeder, V. Drury, and U. Meyer. "Towards Personalized Game-Based Learning in Anti-Phishing Education". In: *2020 IEEE 20th International Conference on Advanced Learning Technologies*. International Conference on Advanced Learning Technologies. Ed. by M. Chang et al. ICALT '20. Tartu, Estonia: IEEE, 2020, pp. 65–66. DOI: 10.1109/ICALT49669.2020.00026.

[G] R. Roepke, V. Drury, U. Meyer, and U. Schroeder. "Exploring Different Game Mechanics for Anti-phishing Learning Games". In: *Games and Learning Alliance*. Ed. by F. de Rosa, I. Marfisi Schottman, J. Baalsrud Hauge, F. Bellotti, P. Dondio, and M. Romero. GALA '21. Cham: Springer International Publishing, 2021, pp. 34–43. DOI: 10.1007/978-3-030-92182-8_4.

[H] R. Roepke, V. Drury, U. Schroeder, and U. Meyer. "A Modular Architecture for Personalized Learning Content in Anti-Phishing Learning Games". In: *SE-SE 2021 : Software Engineering 2021 Satellite Events - Workshops and Tools & Demos*. Software Engineering 2021 Satellite Events. Ed. by S. Götz, L. Linsbauer, I. Schaefer, and A. Wortmann. Vol. 2814. SE-SE '21. Braunschweig, Germany: CEUR, 2021, pp. 1–8. DOI: 10.18154/RWTH-2021-02420.

[I] R. Roepke, V. Drury, U. Meyer, and U. Schroeder. "Better the Phish You Know: Evaluating Personalization in Anti-Phishing Learning Games". In: *Computer Supported Education*. Ed. by H. C. Lane, S. Zvacek, and J. Uhomoibhi. Vol. 1220.

CSEDU '22. Cham: Springer International Publishing, 2022, pp. 71–87. DOI: 10.5220/0011042100003182.

[J] R. Roepke, V. Drury, U. Meyer, and U. Schroeder. "Exploring and Evaluating Different Game Mechanics for Anti-Phishing Learning Games". In: *International Journal of Serious Games* 9.3 (2022), pp. 23–41. DOI: 10.17083/ijsg.v9i3.501.

[K] R. Roepke, V. Drury, P. Peess, T. Johnen, U. Meyer, and U. Schroeder. "More Than Meets the Eye - An Anti-Phishing Learning Game with a Focus on Phishing Emails". In: *Games and Learning Alliance*. Ed. by K. Kiili, K. Antti, F. de Rosa, M. Dindar, M. Kickmeier-Rust, and F. Bellotti. Vol. 13647. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 118–126. DOI: 10.1007/978-3-031-22124-8_12.

[L] R. Röpke, K. Larisch, S. Schöbel, and U. Schroeder. "Mit der Lupe unterwegs - eine spiel-basierte Lernanwendung zu Sicherheit im Internet". In: *DELFI 2019*. Die 17. Fachtagung Bildungstechnologien. Ed. by N. Pinkwart and J. Konert. DELFI '19. Bonn: Gesellschaft für Informatik e.V, 2019, pp. 315–316. DOI: 10.18420/delfi2019_217.

[M] R. Röpke. "Spielbasierte Lernanwendungen für sicheren Umgang mit IT-Systemen und dem Internet". In: *Beiträge der Doktorandenkolloquiums zur DELFI 2019*. Die 17. Fachtagung Bildungstechnologien. Ed. by R. Zender. DELFI '19. Bonn: Gesellschaft für Informatik e.V., 2019, pp. 1–6.

[N] R. Röpke. "Spielerisch sichere Teilhabe: Ein Review spiel-basierter Lernanwendungen über IT-Sicherheit und Sicherheitspraktiken". In: *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung* 36 (2019). Ed. by T. Riplinger, J. Hellriegel, and R. Bolten, pp. 170–185. DOI: 10.21240/mpaed/36/2019.11.21.X.

[O] R. Roepke and U. Schroeder. "Teaching Defence Against the Dark Arts Using Game-Based Learning: A Review of Learning Games for Cybersecurity Education". In: *Computer Supported Education*. Ed. by H. C. Lane, S. Zvacek, and J. Uhomoibhi. Vol. 1220. CSEDU '19. Cham: Springer International Publishing, 2019, pp. 71–87. DOI: 10.1007/978-3-030-58459-7_4.

[P] 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: *Proceedings of the 11th International Conference on Computer Supported Education*. International Conference on Computer Supported Education. Ed. by H. Lane, S. Zvacek, and J. Uhomoibhi. Vol. 2. CSEDU '19. Heraklion, Crete, Greece: SciTePress, 2019, pp. 58–66. DOI: 10.5220/0007706100580066.

[Q] R. Röpke and U. Schroeder. "Mit G/P/S durch die Welt der spielbasierten Lernanwendungen und Serious Games für IT-Sicherheit". In: *DELFI 2019*. Die 17. Fachtagung Bildungstechnologien. Ed. by N. Pinkwart and J. Konert. DELFI '19. Bonn: Gesellschaft für Informatik e.V, 2019, pp. 317–318. DOI: 10.18420/delfi2019_295.

[R] S. Schoebel, R. Roepke, and U. Schroeder. "Phishing Academy: Evaluation of a Digital Educational Game on URLs and Phishing". In: *Games and Learning Alliance*. Ed. by F. de Rosa, I. Marfisi Schottman, J. Baalsrud Hauge, F. Bellotti, P. Dondio, and M. Romero. GALA '21. Cham: Springer International Publishing, 2021, pp. 44–53. DOI: 10.1007/978-3-030-92182-8_5.

Supervised Thesis Projects

In the context of this dissertation, I supervised different student theses, whose results were partially integrated in an adapted form. It is indicated in the text in which parts and to what extent the students' work was utilized in this dissertation.

[1] D. Bayrak. "Implementierung eines interaktiven E-Mail-Interface für Anti-Phishing-Lernspiele". BA thesis. Aachen: RWTH Aachen University, 2020. 67 pp. DOI: 10.18154/RWTH-2021-00081.

[2] A. Brüggemann. "Entwurf und Implementierung eines Lernspiels über die Verwendung von digitalen Zertifikaten im World Wide Web". BA thesis. Aachen: RWTH Aachen University, 2019. 75 pp.

[3] R. deGroot. "Entwicklung einer Remote-Teacher Komponente für das MTLG-Framework". BA thesis. Aachen: FH Aachen, 2018. 68 pp.

[4] M. Drevers. "Design and Implementation of a Digital Learning Game about Email Phishing". MA thesis. Aachen: RWTH Aachen University, 2019. 77 pp.

[5] F. Fernholz. "Konzeption und Entwicklung eines Lernspiels zur Erstellung von Phishing-URLs". BA thesis. Aachen: RWTH Aachen University, 2019. 49 pp.

[6] T. Johnen. "Collection of User Data for Personalized Phishing Education". BA thesis. Aachen: RWTH Aachen University, 2020. 44 pp.

[7] K. Kämmerling. "Personalisierungsschnittstelle für Anti-Phishing-Lernspiele". BA thesis. Aachen: RWTH Aachen University, 2020. 34 pp. DOI: 10.18154/RWTH-2022-00804.

[8] P. Peeß. "Konzeption und Implementation eines Level-Generators für personalisierte Anti-Phishing-Lernspiele". BA thesis. Aachen: RWTH Aachen University, 2020. 71 pp.

[9] S. Schöbel. "Phishing Academy: Entwicklung und Umsetzung eines digitalen Lernspiels zu Website-URLs und Phishing". BA thesis. Aachen: RWTH Aachen University, 2019. 65 pp. DOI: 10.18154/RWTH-2019-02326.

[10] S. Schöbel. "Erweiterung und Evaluation des digitalen Lernspiels Phishing Academy". MA thesis. Aachen: RWTH Aachen University, 2021. 114 pp. DOI: 10.18154/RWTH-2021-08887.

[11] L. Sorressa. "Konzeption eines Lernspieles zur Klassifizierung von Phishing-URLs". BA thesis. Aachen: RWTH Aachen University, 2019. 57 pp. DOI: 10.18154/RWTH-2019-08832.

[12] S. Timpe. "Konzeption und Implementierung von Dynamic Difficulty Adjustment in Anti-Phishing-Lernspielen". BA thesis. Aachen: RWTH Aachen University, 2021. 144 pp. DOI: 10.18154/RWTH-2022-00802.

Contents

1 Introduction

This chapter motivates the topic of this dissertation by providing necessary context information and outlining the shortcomings of existing anti-phishing learning games. Based on this introduction, the main research question and four sub research questions are derived that guide the work presented in this dissertation. Lastly, a brief outline of the dissertation is described to foreshadow the content of different chapters.

1.1 Motivation

Phishing as a threat

One threat to Internet users worldwide is *phishing*, "a scalable act of deception whereby impersonation is used to obtain information from a target" [Las14, p. 8]. The objectives of phishing may include financial gain, identity theft, corporate espionage, and malware distribution [YNT08]. It presents a threat to private end-users but also organizations and businesses, as phishing can target their employees as well [AL13; Sch15]. Thus, phishing can cause financial damage, data loss as well as personal or corporate reputational damage. Recent trend reports summarize the threat landscape and state high numbers of newly created phishing websites [APW21] and clicks on phishing links [Kas21]. While the attackers, so-called *phishers*, employ a diverse repertoire of attack vectors, including email, instant messaging, and even voice phishing [AZ17; YA14], these trend reports indicate that phishing attacks present an imminent threat to users.

Technical countermeasures to phishing attacks

Researchers have explored various measures to fight the threat from different, often interacting directions. The most commonly used technical approach uses blocklists that contain known phishing URLs (Uniform Resource Locators) and are directly integrated into many popular browsers [Oes+19]. While blocklists offer low false positive rates and high explainability, the manual submission and review process required for their creation and maintenance leaves users vulnerable for a short amount of time [She+09; Oes+20]. To this end, machine learning approaches can be applied to classify new websites on demand [Das+20] and thus shorten the 'window of opportunity' [She+09]. Other technical measures include strong authentication, such as Universal Second Factor (U2F) devices which may prevent many types of phishing attacks but are not widely accepted by users [DDC18].

Human factor in phishing attacks

Since technical countermeasures currently fail to stop the threat completely, researchers have also studied the human factor in phishing attacks. Here the questions arise why users are susceptible to phishing and which knowledge and skills users need to recognize and prevent phishing attacks. Research shows that users fail to recognize indicators, e.g., when classifying websites [DTH06] they focus more on the website content than the website URL. Other studies also include psychological aspects of phishing perception, e.g., using the principles of persuasion [Cia06]. A recent study on the users' reading abilities shows differences in classifying different URL categories, e.g., long subdomains are the most confusing category of URLs [Rey+20]. Overall, by studying the human factor in phishing attacks, two requirements for users to prevent phishing can

be identified: (1) users need to know what to look for to detect phishing attacks, and (2) they need the situational awareness to apply it at the right moment.

User education

User education has been explored as a complementary approach to technical countermeasures to raise awareness and teach the necessary knowledge and skills to detect and prevent phishing attacks [KS12]. *Anti-phishing education* is a research field in the domain of security education. Security education is a cover term for concepts and approaches of information security awareness, training, and education. The objective of anti-phishing education is to educate users about phishing by raising awareness and teaching the relevant knowledge and skills to detect phishing attacks and apply appropriate countermeasures. Various institutions and actors have provided different materials and approaches on the topic, e.g., financial institutions [Spa; Pay; N26], federal institutions [Bunb; Buna; Fed], and research groups [SEC21]. Learning resources range from textual information and video-based content [Buna] to interactive phishing quizzes for users to test their knowledge [SEC21]. Common approaches in anti-phishing education also include awareness campaigns using simulated phishing attacks [VSB20].

Anti-phishing learning games

Similarly, researchers have explored game-based learning for anti-phishing education by developing so-called *anti-phishing learning games*. The key idea of anti-phishing learning games is to engage users in learning about phishing by providing an interactive learning environment. The game environment presents the opportunity to apply learned knowledge and practice skills regarding the detection of phishing attacks. A common theme in anti-phishing learning games requires learners (or players) to classify URLs or emails as either legitimate or phishing. The games may use digital storytelling to guide learners and introduce relevant knowledge, e.g., about different types of phishing URLs. Next, the conveyed knowledge needs to be applied when learners are required to classify URLs or emails in the game. This way, skills to recognize indicators for phishing are practiced. Depending on the game and the context it is used in, anti-phishing learning games can be used to raise awareness, teach necessary knowledge and provide practice opportunities free of risks and consequences. The effectiveness of existing games has been evaluated in different user studies (e.g., [She+07; BC14; ALM15; Wen+19]). While the games can be designed for particular target groups, specific educational contexts, and with a requirement of prior knowledge in Computer Science (CS) or Information Technology (IT) security, they can also be kept rather generic and without any requirements for prior knowledge, thus, suitable for end-users in informal learning contexts, e.g., the general public.

Problems with unknown services

While the game mechanic of classifying URLs or emails as legitimate or phishing does mimic the real-world actions learners should be able to perform, it also has its weaknesses, especially in the context of educational games. To this end, learners are required to apply their knowledge and distinguish between trustworthy and malicious services or origins (i.e., the company which provides the content of a website or is the sender of an email, e.g., the company 'PayPal' for the website 'https://www.paypal.com'). Although the learning game may teach systematic rules to determine whether a URL or email is part of a phishing attack, learners might be unable to decide for URLs and emails of unknown services due to a lack of reference. If indicators are not conclusive, not knowing the service presented in the game requires learners to guess, e.g., they do not know the original URL of the service and cannot decide whether a given URL is legitimate or not. Instead, learners should actively discard the URL as it is unknown. The inability to decide may weaken the game's potential for practice, since learners cannot relate between correct classifications and applying previously learned knowledge about indicators. Furthermore, unknown services can hamper raising awareness since it might be harder for learners to relate to the game's content and transfer learned knowledge

to their daily activities. Both might negatively impact the learning experience and the learners' abilities in detecting phishing attacks.

Limitations in game design

Further limitations may apply when games implement the common decision mechanic in which learners have to classify URLs as legitimate or phishing. By relying on a binary decision mechanic, the possibility of advancing in the game through guessing is increased, and the games' assessment capabilities are limited. As such, incorrect decisions do not reveal any possible misconceptions or problems learners might have and conveyed knowledge or skills are not evaluated. To this end, there is an untapped potential for such games, and a redesign exploring different game mechanics could allow more detailed assessment and feedback to guide learners and support the learning process.

Personalization of anti-phishing learning games

To solve the problem with classifying unknown services and provide more detailed assessment and feedback, it is suggested to design new anti-phishing learning game prototypes exploring different game mechanics. Further, the objective is to personalize anti-phishing learning games and consider learner characteristics, e.g., learners' familiarity with different services. To the best of the author's knowledge, there is no existing work on personalized anti-phishing learning games. There is a lack of systematic approaches to generate learning game content for anti-phishing learning games dynamically, and no personalized anti-phishing learning games have yet been studied regarding their effectiveness. This reveals a research gap and raises the questions of how personalized learning games can be utilized to support learners in detecting phishing attacks and how to design a framework for the personalization of anti-phishing learning games.

Focus and target group

For the scope of this work, the focus is on anti-phishing learning games about phishing URLs. The selected target group is end-users with little to no prior knowledge of IT security. The topic and target group were selected to support the creation of comparable game prototypes with specific learning goals and detailed learning content that could be evaluated with an accessible user group and no additional resources to account for specific prior knowledge.

Conclusion

To summarize, phishing is still an imminent threat to Internet users worldwide, although there is significant research to stop phishing by various means, e.g., technical countermeasures and understanding the human factor in phishing attacks. Anti-phishing education presents a complementary approach to fighting phishing and as such, researchers have explored game-based learning by developing various anti-phishing learning games to teach learners how to detect and protect against phishing attacks. While some of these games have proven to be effective in different studies, limitations in their design may raise the question of how personalization and different approaches to game design could be used to support learners with detailed assessment and feedback and improve the learning experience. The work presented in this dissertation is guided by this motivation and contributes to the research domain of personalized anti-phishing learning games.

1.2 Research Questions and Contributions

Based on the motivation described in the previous section, the main research question (MRQ) for this dissertation reads as follows:

Main research question

How can personalized anti-phishing learning games be utilized to support end-users with little to no prior knowledge of IT security in detecting phishing URLs?

To answer the MRQ systematically, the following set of sub research questions (SRQs) will be answered throughout this dissertation:

Sub research question 1

SRQ-1: How do existing anti-phishing learning games consider characteristics of individual learners using personalization?

For the first SRQ, the focus is on generating an overview of existing anti-phishing learning games and evaluating whether personalization could be integrated to support individual learners. As such, the objective is to systematically review existing literature and available implementations to provide an overview on existing anti-phishing learning games and identify potential research gaps.

Sub research question 2

SRQ-2: How can the design of existing anti-phishing learning games be improved to support learners in detecting phishing URLs?

The focus of the second SRQ is on the evaluation of the game design of existing anti-phishing learning games. Further, it focuses on the deduction of design implications for developing new game prototypes that enrich the state of the art and allow for personalization. Thus, the objective is to identify drawbacks of current game designs, provide new game prototypes implementing a different game design, and prepare for the personalization of learning game content.

Sub research question 3

SRQ-3: How can personalization be integrated into anti-phishing learning games to consider learner characteristics and support the learning process?

For the third SRQ, the main focus is on identifying a systematic approach to personalization in anti-phishing learning games. The objective is to conceptualize and implement a personalization framework that can provide personalized learning game content for individual learners in the developed game prototypes.

Sub research question 4

SRQ-4: How does the personalization of anti-phishing learning games influence the learners' performance in detecting phishing URLs?

The fourth and last SRQ focuses on evaluating the developed game prototypes, with or without personalization, and gaining insights on their effectiveness and the influence of personalized learning game content on learners' performance in detecting phishing URLs. To this end, the objective is to conduct a comparative user study using the different game prototypes and the personalization framework.

Contributions

By addressing the MRQ and its sub research questions in this dissertation, the following major contributions are provided to the research domain of personalization in game-based anti-phishing education:

- A systematic overview on existing literature and available implementations of learning games for anti-phishing education (as well as security education)
- Two new anti-phishing learning game prototypes utilizing different game mechanics than previous games in order to allow more insights into the learners' actions

and decision processes as well as suitable interfaces for the personalization of learning game content

- A framework for personalizing anti-phishing games using three distinct modules for (1) data collection and learner modeling, (2) content generation, and (3) content delivery
- An evaluation comparing the different game prototypes with or without personalization and insights regarding the influence of personalized learning game content on learners' performance in detecting phishing URLs

1.3 Dissertation Outline

This dissertation is organized into ten chapters, whose content is briefly introduced here.

Chapter 1 introduces the reader to the topic of the dissertation and provides a problem statement as well as the research questions, objectives, and contributions.

Chapter 2 presents the scientific foundations and terminology which are important throughout this dissertation. It introduces concepts and terms from three different domains: (i) security education, training, and awareness, (ii) game-based learning, and (iii) personalization. The chapter covers the important definitions and objectives of each domain. Furthermore, it covers domain-specific foundations to provide context to this work.

Chapter 3 contains a brief presentation of related work and the current state of research and technology. It presents important work in the field of personalization in learning games, security education, and learning in general, since to the best of the author's knowledge no immediate related work in the field of personalized anti-phishing learning games exists. Furthermore, the chapter introduces research gaps regarding the personalization of anti-phishing learning games.

Chapter 4 presents the combined research design of design-based research and experimental hypothesis-testing research applied in this dissertation and elaborates on the different methods and tools used to answer the research questions of this dissertation.

Chapter 5 continues the presentation of related work by describing the methodology and results of two systematic literature reviews on learning games for security education and anti-phishing education. This chapter presents existing learning games as well as various analyses classifying available games in aspects like target group, educational context, learning goals and learning content. The results provide additional research gaps regarding the design of existing anti-phishing learning games and serve as a starting point for the development of new game prototypes.

Chapter 6 presents the design and implementation of two new anti-phishing learning game prototypes utilizing different game mechanics and thus providing more insights into learners' actions and decisions. Furthermore, an additional implementation is presented that serves as a baseline in the comparative user study.

Chapter 7 introduces the concept and implementation of the personalization pipeline, a framework for providing personalized learning game content in anti-phishing learning games. The personalization is based on the learners' familiarity with different services when creating game content like URLs or emails.

Chapter 8 presents two comparative user studies and a longitudinal study using the different game prototypes and the personalization pipeline. The evaluation provides insights into the differences between different game mechanics as well as between personalized and non-personalized versions of a game. In addition, the longitudinal study provides results regarding the long-term effects and self-reported behavioral changes of the different game prototypes.

Chapter 9 summarizes the findings and results of the research process. Lessons learned from the evaluation of different game prototypes, and the use of personalization are discussed. Further, the contributions in response to this dissertation's research questions and objectives are described.

At last, Chapter 10 concludes the research and practical work in this dissertation, summarizes the contributions, and outlines future work opportunities and challenges in the field of personalized game-based anti-phishing education.

2 Foundations

This chapter introduces the relevant concepts, terms, theories, and the important aspects of different research areas closely related to the topic of this dissertation. Each aspect is presented and observed from the perspective of personalizing game-based learning for security education, particularly anti-phishing education.

2.1 Security and Anti-Phishing Education

Due to the topic of this dissertation being in the domain of security education, related concepts and terms are explained in this section. It presents an overview of important definitions and objectives, as well as the domain's scopes, target groups, and methods. Further, phishing as a topic of security education is reviewed and relevant concepts and terms for anti-phishing education are introduced.

2.1.1 Definitions

Information security

At first, a definition of *Information security* is needed to set a basis for related terms and concepts. According to the ISO/IEC 27000:2018 standard, information security is defined as the "preservation of confidentiality, integrity, and availability of information" [ISO18, p. 4]. Another definition by the Committee on National Security Systems describes it as "the protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity and availability" [CNS15, p. 64]. While definitions of different researchers and organizations may vary [ALK14], they most often evolve around three information security attributes or qualities: *Confidentiality*, *Integrity*, and *Availability*, or the *CIA-triad*. They are also called primary security goals. While the origin of the CIA-triad is challenging to determine, it presents as an essential concept for information security professionals for more than two decades [Fru20]. In literature, further security goals are mentioned like authenticity, accountability, non-repudiation, and reliability [ISO18].

IT security

The concept of *IT security* is a part of information security that focuses on the protection of sociotechnical systems [Eck12]. These systems have two inseparable components: the social component (i.e., one or more humans) and the technical component (i.e., one or more systems). Thus, IT security addresses the protection of systems and their users [Eck12]. In literature, IT security may also be addressed by using the synonymous term *cyber security*.

Awareness, training, and education

Similar to information security and IT security, the terms information security awareness, information security training, and information security education are often defined differently in different contexts. They are sometimes even used interchangeably. In [ALK14], these terms are untangled by reviewing existing definitions and analyzing them with a focus on three key aspects: (1) *focus*, (2) *purpose*, and (3) *methods*. The

analysis shows that at least one of the key aspects is often missing in different definitions. Lastly, they present new definitions containing all three key aspects to solve the confusion. It should be noted that the presented definitions are from an organizational perspective:

- *Information security awareness* (ISA) denotes "any endeavor to focus employees' attention on information security in order to ensure that all employees understand their roles and responsibilities in protecting the information that is in their possession" [ALK14, p. 250]. This can be achieved using different types of media.
- *Information security training* (IST) denotes "any endeavor that is undertaken to ensure that every [end-user] is equipped with the information security skills and information security knowledge specific to their roles and responsibilities by using practical instructional methods such as seminars and workshops" [ALK14, p. 250].
- *Information security education* (ISE) denotes the "endeavor to provide insight into and an understanding of information security documents in order to ensure that every employee is equipped with necessary information security skills and information/knowledge to protect specific information" [ALK14, p. 249]. This can be achieved using academic or formal instructional methods.

While many definitions for ISA, IST and ISE focus on the organizational context, they can be adapted to the more informal learning context and the target group of end-users. For the scope of this dissertation, these definitions can be adapted by simply replacing the word 'employee' with 'end-user'. Besides, methods mentioned in the definitions should be understood as examples or categories.

Both, IT security and information security, are concepts defining the terms of security education, training, and awareness. While they evolve around employees and their roles and responsibilities, awareness is the first or lowest level in the program and addresses the largest user group, i.e., all users. It serves as a catalyst to the training part and should provide motivational information to learn [Hig05]. Its goal is to create a security awareness culture among all users of an organization. With specific roles and increasing responsibilities, security training and education are needed. When translating it to the context of general end-users in informal learning contexts, the change of roles and increase of responsibilities cannot be denoted by job titles or task descriptions. It might be harder to differentiate whether end-users require awareness, training, or education.

Security literacy

Another related concept is *security literacy*. While the term is used in different publications [FM14; JI16], a proper definition is often missing. It is introduced in the IT security learning continuum as part of a transition between awareness and training [Wil+98] and in its most recent revision, security literacy is defined as "an individual's familiarity with a basic set of knowledge with security principles and practices" [TK14, p. 31]. However, it presents an additional concept besides security awareness, training and education, which is mainly focused on the basic knowledge of security principles and practices. It therefore increases the potential ambiguity and is not considered further in the scope of this work.

In conclusion, the terms security awareness, training, and education need to be defined for a specific target group and context. In the literature, the different terms are often used interchangeably or without clear definitions, thus, increasing the potential for ambiguity, confusion and misconceptions. In the scope of this dissertation, the term *security education* is used to reduce the potential and provide a common wording throughout the research. As such, security education is defined as the endeavor to provide awareness and understanding of security concepts to ensure that every end-user

is equipped with the necessary knowledge and skills to protect themselves and personal data depending on their interactions with IT systems and the Internet. It can be achieved using different instructional methods, e.g., learning games. The definition is an adaptation and combination of definitions presented in [ALK14]. Since this dissertation focuses on anti-phishing education, a particular subdomain of security education, it is separately introduced in Chapter 2.1.3.

2.1.2 Objectives

In the following, the objectives and reasons for security education will be discussed. Thereby, it will be distinguished between three different contexts: (1) organizational, (2) educational, and (3) private.

Objectives in the organizational context

In organizational contexts, security education is one measure to ensure the protection of the organization, minimize risks and reduce the number of security incidents due to the lack of employees' awareness [Hig05]. While there are different (technical) security measures in place (e.g., firewalls, network monitoring, anti-virus software), security education targets the employees of an organization. It addresses the human factor of security, as employees need to behave adequately to ensure that even secure systems are not threatened. As described above, security education differs for different groups of employees with different roles and increasing responsibilities [Cab17; TK14]. The scope is aligned with the employees' tasks and position within the organizational structure. While all employees should be part of security awareness programs, employees working with IT systems and sensitive information might be required to participate in security training and education programs. This way, different types of interventions are also bound to the levels of IT proficiency. Security education programs should establish a well-rounded security attitude or culture across the whole organization [Cab17] and enable all employees to understand basic security terms and concepts through establishing security literacy [TK14]. Especially when systems fail, the educated employees are "the last line of defense in a network" [Hig05, p. 1].

Objectives in the educational context

In educational contexts, e.g., higher education, the objectives of security education may differ depending on the user group. Security education is part of a curriculum and aims at educating students about the theories, concepts, and applications of IT security. Topics vary but may include cryptography, system security, and more. The subject matter is more theoretical or applied depending on the course type. Its purpose is to educate the future workforce and train IT security professionals for industry or research. Especially students of IT security study programs focus on theoretical and research-oriented aspects of IT security. However, in other CS-related study programs, the portion of security education can vary based on the course choices. Lastly, students of other study programs (not related to IT security or CS) can encounter security education from different perspectives, e.g., through philosophical or business-oriented perspectives. While in educational institutions further target groups can be considered (e.g., instructors, researchers), they can all be categorized as employees in an organization. For the scope of this work, both angles on security education in the educational context are not discussed further, as the focus lies on end-users in private contexts.

Objectives in the private or personal context

Compared to organizational and educational contexts, security education for end-users in their personal, day-to-day activities aims at building a security-aware attitude and teaching necessary knowledge and skills to protect themselves and others. According to [BS14], the primary purpose of security awareness campaigns is to influence the adoption of secure behavior online. Security education also provides a complementary

method of supporting users in staying safe online. Although many technical security measures can be implemented, they are considered as difficult to understand and require configuration. Therefore, security education is used as an additional, complementary approach. however, outside of organizational structures or formal educational contexts, the level of IT proficiency might not be as easy to determine. While many end-users may gain a low but possibly sufficient level of IT proficiency through experience from their day-to-day activities online and when using IT systems, employees and students are also end-users in private contexts. Thus, IT proficiency can be potentially higher. Therefore, the target group of end-users in private contexts can be divided into subgroups by determining their prior knowledge of CS and experience with IT systems. For end-users with advanced prior knowledge in CS or IT security, security education can quickly become boring and repetitive. While this group of end-users might enjoy more specialized educational interventions, end-users with little to no prior knowledge of CS or IT security need to start from the beginning. In this case, the questions arise where to begin and how to set the objectives. Therefore, security education tries to ensure that every end-user has the necessary skills and knowledge to protect themselves in their interaction with IT systems and the Internet by increasing awareness and supporting the understanding of security concepts and topics (as defined in Chapter 2.1.1).

Conclusively, security education aims to enable end-users to behave securely and protect themselves and others, whether in private day-to-day activities, educational contexts, or on the job. For the scope of this work, the objectives of security education are to build a security-aware attitude and teach the necessary knowledge and skills to act securely in any context. It does not include domain-specific aspects (e.g., organizational structures) but focuses on teaching CS-related principles, concepts, and practices. It should be noted that content-specific objectives, i.e., learning objectives of particular topics in security education are not discussed here. Instead, the objectives of anti-phishing education are introduced in Chapter 2.1.3. The focus is on security education for private end-users with little to no prior knowledge in CS or IT security. Both target group and scope should be considered during the design, development, and evaluation of security education interventions. However, the outcomes and results do not have to be limited and can be discussed for the other scopes and target groups introduced above.

2.1.3 Anti-Phishing Education

One particular field of security education covers the threat of phishing and possible countermeasures: *Anti-phishing education*. In this section, necessary definitions and concepts in the field of anti-phishing education are introduced.

Definition and origin of phishing

The term *phishing* is widely used in various contexts (e.g., scientific literature, media and by organizations), however, no consensual definition exists [Las14; KIJ13]. This makes the concept of phishing challenging to grasp and can lead to confusion. [Las14] presents a systematic literature review and synthesis of a new definition based on existing ones. The result is a rather abstract definition of phishing as "a scalable act of deception whereby impersonation is used to obtain information from a target" [Las14, p. 8]. While there is a lack of consensus regarding phishing definitions, the word's origin presents more clearly. The word was first used around 1995-1996 [KIJ13] and is a leetspeak variant of the word 'fishing', where the *f* is replaced by *ph* [Mey06]. Similar to fishing, in phishing, attackers use bait (e.g., deceptive messages) to catch a fish (e.g., steal confidential information of victims) [KIJ13].

Objectives of phishing

The primary objectives of phishing attacks are financial gain, identity theft, corporate espionage, and malware distribution [YNT08]. Phishing attacks target the weakest link of even secure systems: the user [AL13; Sch15]. The associated damages include, but are not limited to:

- financial damage or loss,
- data loss, and
- personal or corporate reputational damage.

According to the Anti-Phishing Working Group (APWG), the number of phishing attacks doubled in 2020, with a record of 225,304 phishing websites in October. Most targeted were financial institutions, webmail-based or software-as-a-service organizations, and payment providers [APW21]. Regularly published reports by the APWG show that even after 25 years, phishing is still a threat today.

Social engineering

A key deception technique in the context of phishing is *social engineering*. Similar to the concept of phishing, social engineering shares various definitions without having a consensus within the research community. Wang et al. summarize the evolution of the concept and its definitions and propose a new consensual definition tailored to the domain of IT security [WSZ20]. They define social engineering as "a type of attack wherein the attacker(s) exploit human vulnerabilities by means of social interaction to breach cyber security, with or without the use of technical means and technical vulnerabilities" [WSZ20, p. 85105].

Channels and types of phishing

Phishing attacks utilize different scalable channels to deceive users into divulging information. In generic phishing attacks, messages are often distributed in bulk, i.e., mass distribution to large user groups. Common channels include emails, social networks, phone calls, SMS (Short Message Service), or comparable messaging services (e.g., WhatsApp, Telegram). Based on the selected communication channel, different types of phishing are mentioned in the literature, e.g., email phishing, vishing (i.e., voice phishing attack whereby a phone call by the attacker is used to lure victims into divulging information [YA14]), smishing (i.e., phishing based on SMS or text messages [YA14]). Compared to generic phishing attacks, spear-phishing is a specific type of phishing that targets individuals or organizations using personalized, tailored attacks (e.g., personalized deceptive messages).

Description of a typical phishing attack

In order to make the described concepts more accessible, the following description introduces the typical phishing attack [Oes+18]: First, the phisher (i.e., the attacker) selects a website and spoofs it by copying its look and feel such that it is difficult to distinguish between it and the legitimate counterpart. Next, the attacker sends a message to the targeted user, leveraging deception through social engineering to lure the user into clicking on a URL to the phishing site. If the user clicks on the URL to visit the phishing website and submits the desired information, the phishing website transmits this information to the phisher. Finally, the phisher can misuse it either directly or indirectly (e.g., selling it to gain profit).

Technical countermeasures

In the effort of fighting the phishing threat, research encompasses various countermeasures from different, often interacting directions. Among technical countermeasures, the most commonly used are blocklists that contain known phishing URLs and are directly integrated into many popular browsers [Oes+19]. While these lists offer low false positive rates and high explainability, users are vulnerable for a short amount of time due to manual submission and review processes in the creation and maintenance of blocklists [She+09; Oes+20]. To overcome these issues, different data science-based

approaches (e.g., machine learning, data mining) have gained popularity among researchers [Das+20]. This allows quicker blocklist maintenance or even real-time classification when users open a website or email. A more user-centered approach to prevent phishing focuses on strong authentication, e.g., U2F devices prevent many types of phishing attacks but are not widely deployed or accepted by users [DDC18]. In the end, technical countermeasures and phishing attacks are in a constant cat-and-mouse game: When new technical solutions arise, phishers adapt their approach to circumvent existing mitigations and vice versa, resulting in technical solutions alone failing to stop phishing.

User education

Since technical countermeasures do not stop the threat entirely, and users are still vulnerable, researchers turn to *user education* [KS12] as a complementary approach. By educating end-users about the threat, useful indicators, potential damages, and possible measures to prevent phishing, end-users can take action in protecting themselves. At first, however, the question arises, why users are susceptible to phishing in the first place. In [DTH06], a study is presented, where they asked participants to decide whether 20 different websites were legitimate or not [DTH06]. As a result, only a few participants actually used indicators in the browser and focused instead on website content [DTH06]. The accuracies in classifying websites ranged between 9% and 100% [DTH06]. More recent studies on the URL reading ability of end-users found that identifying the URL destination is difficult, and strategies are inconsistent due to the complexity of the URL structure and misleading, deceptive keyword placement within the URL [Rey+20; AVW20]. Furthermore, [GWD17] presents a review of various publications on susceptibility to phishing [GWD17] and summarizes different factors, e.g., cognitive limitations, emotional arousal, and personality traits. Other studies also include psychological aspects, e.g., using the principles of persuasion to deceive end-users [Cia06]. In conclusion, human vulnerability is due to the susceptibility to deception [GWD17].

Educational resources and interventions

Different institutions and actors in the field of anti-phishing education have provided different materials and approaches on the topic, e.g., financial institutions and services [Spa; Pay; N26], government institutions [Bunb; Buna; Fed] and research groups [SEC21]. Besides textual information, some resources also offer video-based content [Buna; SEC21] and interactive phishing quizzes [SEC21]. A common approaches in anti-phishing education is awareness campaigns, in which end-users receive some kind of training and are afterwards tested using simulated phishing attacks [VSB20]. Another type of intervention relies on game-based learning, where researchers have developed and studied games as a suitable educational intervention. For details on existing anti-phishing learning games, relevant related work, and extensive systematic literature reviews are presented in Chapter 3 and Chapter 5.

Learning content

While the before-mentioned resources and interventions may approach the topic of phishing in different ways, they usually present a definition and possible examples for phishing attacks. They also emphasize different indicators on recognizing phishing in emails or different communication channels. As the threat spreads across different channels, information materials vary between the depth and coverage of different phishing attacks. A commonality between many educational resources or interventions is the focus on malicious, manipulated URLs [KS12; Ara12] as the URL is one of the most reliable features in phishing detection [Vol+17; Gar+07].

In [FA19], it is argued that since phishers adapt quickly to discovered obfuscation techniques, possible educational interventions might be outdated quickly. As most phishing attacks leverage users to access a phishing website by clicking on a URL, the URL is a possible starting point to educate the user. This means educating users about the basic

URL structure, how to read URLs, and possible distinctions between malicious and benign URLs. *URL parsing* denotes the skill of reading a URL and identifying its components (e.g., scheme, top-level domain, path). As research shows, users have difficulties in identifying the destination of clearly written URLs [AVW20; Rey+20]. Various manipulation techniques make recognizing phishing URLs even harder if the basic URL structure is not identified correctly. URL parsing can, therefore, be a valuable skill in order to divide a URL into its components and detect possible manipulations. A *URL manipulation* describes an alteration, addition, omission, or other types of manipulation of one or more components of a URL. In the case of phishing, the intention is to deceive a user by masking a malicious URL as trustworthy and benign-looking such that users cannot distinguish it from an original URL. Manipulation techniques include typo-squatting (i.e., mistyping popular domain names [Agt+15]), combo-squatting (i.e., combining popular domain names with one or more phrases [Kin+17]), and subdomain/path embedding (i.e., embedding popular target domains into the subdomain or path of a URL [Rob+19]). Further details on different manipulation techniques, including examples, are presented in Chapter 6.

In conclusion, anti-phishing education is a subgenre of security education that deals with the threat of phishing and possible countermeasures. As many phishing attacks utilize URLs to lure users to malicious websites, a particular approach of anti-phishing education teaches users to parse URLs and recognize different manipulation techniques. It should be noted that this approach is limited to URL-based phishing attacks and might not enable users to recognize and mitigate other phishing attacks. Furthermore, it is important to acknowledge that anti-phishing education should also raise awareness and support risk-aware online behavior to be able to detect phishing and other related threats.

2.2 Game-Based Learning

This section introduces the key terms and concepts of game-based learning and briefly presents the theoretical foundations. As the concept of learning games is particularly important in the scope of this dissertation, relevant aspects regarding their design and implementation are presented as well.

2.2.1 Definitions

Game-based learning

Over the last years, the educators' and researchers' interest in the use of games for learning inside or outside the classroom has grown [Pre03; Gee07; PHK15]. The concept of *game-based learning* intuitively translates to learning by playing games. According to the Enzyclopedia of the Sciences of Learning, it is defined as "learning that is facilitated by the use of a game" [Whi12, p. 1337]. As such, the difficulty in defining game-based learning lies in defining the term *game* [Whi12].

Games and their defining traits

The term *game* can be defined as "a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome" [STZ04, p. 80]. However, a game is often defined concerning the given context, application, or scenario. The variety of platforms, genres, playing modes, and target groups makes it difficult to agree on one widely accepted definition, although researchers agree on common traits or characteristics of games [May14; MJ10]. For instance, McGonigal defines the term *game* by focusing on four shared defining traits [McG11]:

- a goal,
- rules,
- a feedback system
- voluntary participation.

The identification of those four traits allows for comparison between games and matches common characteristics of effective learning experiences as described in Chapter 2.2.3. While some more traits or characteristics could be considered in the definition of a game, they are often domain-specific features of a set of games. Therefore, they are not considered defining traits of a game.

Game mechanics

An commonly used term to analyze or describe games and how the gameplay works (e.g., what players have to do to win or achieve goals within the game) is the term game mechanics. *game mechanics* can be defined as "methods invoked by agents, designed for interaction with the game state" [Sic08]. As such, the agent is equal to the player (or learner in the context of this work), but can also be a non-player character (NPC) or artificial agent. The methods are actions invoked by players to interact with the game, and they are limited by rules that apply to the game environment. According to [Jär08] and [Sic08], game mechanics are best described using verbs (e.g., jump, move, classify, combine, discard). In the scope of this work, the term game mechanics ist used when analyzing existing games and to describe the design and implementation of new game prototypes (see Chapter 6).

Gamification

Besides game-based learning, there is the concept of *gamification*. It is defined as "the use of game design elements in non-game contexts" [Det+11, p. 10] to foster motivation and engagement. In the educational context, this includes adding game-specific features, often involving a reward system or narrative elements, e.g., badges, rankings, leaderboards, avatars, and incentives like points or stars. As gamification only adds game design elements, the learning process itself remains unchanged. Examples in other domains include reward programs of airlines, supermarkets, and coffee shops, where customers are motivated to spend money and gain a special customer status or coupons for their next purchases. [Pla+20]

Playful learning

Another concept related to gamification and game-based learning is *playful learning*. Playful learning does not require a complete game to redesign the learning task but rather subtly use game design elements and provide a playful learning experience. It is different from gamification as it actually changes the learning task. [Pla+20]

Serious games

Games which are designed with purposes other than pure entertainment are called *serious games* [MC05]. While this is just one definition for the term, existing definitions most often share the aspect of a particular purpose, objective, or goal that is not entertainment-related but rather educational. The first definition dates back to the 1970s, where serious games are defined as "games that have an explicit and carefully thought-out educational purpose and are not intended to be played primarily for amusement" [Abt70, p. 9]. Serious games can be applied in various domains, one of which is education and learning. The so-called *learning games* (or educational games) denote a subgroup of serious games which can be designed for formal and informal learning scenarios [Dör+16]. In Chapter 2.2.4, a more extensive elaboration on learning games is presented, as this dissertation focuses on the design and implementation of learning games for anti-phishing education.

Digital game-based learning

With the increasing use of digital devices and digital media in education, as well as the growing digital gaming industry, the concept of *digital game-based learning*, was

defined as combining computer games and educational content [Pre03]. Today, the term *computer game* is somewhat outdated, and there are games for all kinds of platforms, e.g., mobile phones, tablets, multi-touch tabletops, and virtual reality (VR) headsets. A more suitable term to cover today's variety is *digital games*, i.e. "games that use some kind of computing machinery" [Dör+16, p. 2]. As none of the previously presented concepts, i.e., game-based learning, gamification, and serious games, are limited to non-digital environments, they allow the use of digital game design elements. [Dör+16] showcases this in the definition of serious games as digital games which are created for entertainment and at least one additional characterizing goal (e.g., learning about phishing).

In the scope of this dissertation, a slightly modified version of Prensky's definition of digital game-based learning [Pre03] is used: Digital game-based learning is the combination of digital games with educational content.

2.2.2 Objectives

Domain-specific objectives

The utilization of games and their design elements within the learning context can have different objectives and benefits. As such, the objectives of game-based learning are often domain-specific. For example, while in healthcare, game-based learning is used to promote physical or mental health and improve health conditions, in the corporate domain, game-based learning is used to teach processes or tasks and form attitudes or habits. According to [Dör+16], serious game have so-called characterizing goals (i.e., additional goals besides to entertain). In [WH13], the authors provide examples of competence domains specifying the characterizing goals of serious games, e.g., cognitive, social competencies, or personal competencies.

Motivation and engagement

The most prominent reason why educators might turn to game-based learning is to foster motivation and increase engagement of learners. Researchers argue that since games can engage large user groups in individual and social activities, they are a suitable medium for learning [Pre03; Gee07; Squ13]. In addition, aspects like sensory pleasure (e.g., through visuals and sounds) and storytelling can increase enjoyment of the experience and support interest or curiosity [Dör+16]. Although many researchers argue the motivational benefits of game-based learning, it lacks the empirical evidence to proof it [Whi12]. Furthermore, it is difficult to draw general conclusions about game-based learning since it is applied to many different domains and different target groups [HRH20]. Here, evidence-based research approaches are needed.

Emotional objectives

Closely linked to motivation and engagement is the game's potential to reach learners emotionally. Researchers argue that good gameplay can "evoke challenge, suspense, thrill, relief, empathy with characters, or caring for an environment" [Dör+16, p. 5] and therefore support commitment and investment [Dör+16].

Educational objectives

In [Con+12] and [Boy+16], literature on games for entertainment and for learning was reviewed to provide an overview on empirical evidence of positive outcomes of (digital) games. While in [Con+12], a review of the literature published between January 2004 and February 2009 is provided, the updated work [Boy+16] extends the review by including literature published until February 2014. The joint work reports on 272 papers, providing diverse research regarding different positive outcome dimensions of games. The most frequently occurring outcomes of learning games are knowledge acquisition and content understanding across many different subjects (e.g., STEM, health, social issues). This is often associated with the curricular embedding and common intention of using games to educate players about a particular topic, e.g., IT security or phishing.

Safety, failure, and scalability

Another objective of game-based learning is to provide a safe learning environment. This links to concepts like learning by doing and learning by mistakes. A simulated environment provides a safe space for learners to act and react. Since games are consequence-free environments, learners can experiment and make mistakes. According to [PHK15], game-based learning allows for graceful failure. Instead of thinking of failure as an undesirable outcome, failure can be utilized in the learning process. The lowered consequences in games encourage exploration and risk taking. The outcomes of the game lastly do not influence the real world [Whi12]. This is especially useful in domains like pilot training, military, and security education, where learners face risks and danger. Furthermore, game-based learning can be used as a scalable inexpensive solution in domains where resources are limited or expensive. Finally, similar to simulations and VR, games can provide an interactive and exploratory space [Whi12].

Conclusion

As for this dissertation, the primary objectives of game-based learning are to improve the learning experience, support knowledge acquisition, and engage the learners. For security education, games may provide a safe learning environment in which learners are allowed to experiment and make mistakes. For example, they can click on a phishing URL without the fear of unwanted consequences and unknown risks. It, thus, makes the game an environment to practice and learn from mistakes.

2.2.3 Theoretical Foundation

Different perspectives and theories

According to [PHK15], the theoretical foundations of game-based learning cannot be understood by taking only one perspective but rather by approaching it from different theoretical perspectives of learning. Games are a complex genre of learning environments, and depending on the intentions and objectives, different aspects need to be considered to describe the theoretical foundations of game-based learning. In [PHK15], cognitive, affective, motivational, and sociocultural foundations of game-based learning are discussed to emphasize the complexity. In an extensive meta-review, [KSvK21] takes a broad look at the theoretical foundations of gamification, game-based learning, and serious games. They provide an overview of 118 different theories and models used to design and evaluate gamified interventions. Based on their focus, these theories can be divided into three different groups: (a) affect and motivation, (b) behavior, and (c) learning. While all theories try to explain how gamification achieves its positive outcomes, they originate from various research fields, e.g., cognitive psychology, social psychology, and human-computer interaction. Similar to [PHK15], [KSvK21] shows how diverse the ongoing research can be and how different theoretical foundations can be used when working with gamified interventions.

Learning theories and playfulness

When trying to link game-based learning to traditional learning theories, various examples exist that games can be based on different learning theories. In [PHK15], the examples of Angry Birds, Crayon Physics, and World of Warcraft are discussed to argue that different games can utilize behaviorist, cognitivist and constructivist elements in game design. Thus, a game can be based on one or more learning theories. In [PHK15], it is argued that playfulness, the key characteristic of a game, can be applied no matter which learning theory since it is an enriching orthogonal dimension.

Theory of Gamified Learning

With the intention of linking the research literature of serious games and gamification, Landers developed the *theory of gamified learning* [Lan14]. According to the theory of gamified learning, gamification can affect learning through two different processes. On the one hand, gamification can affect learning via *moderation*, i.e., adding game elements to encourage the desired behavior and increase learning outcomes. On the other

hand, learning can be affected via *mediation*. In this case, gamification "intends to encourage a behavior or attitude that will itself improve learning outcomes" [Lan14, p. 763]. In comparison, Landers points out that serious games are different since they are designed to provide instructional content and fulfill the role of instructors. Gamification rather augments or supports pre-existing instructional content by encouraging a behavior change. Although the theory of gamified learning addresses the theoretical foundations of gamification, it is crucial to consider both serious games and gamification as complementary approaches. Both utilize the same game elements toolkit but apply it differently to learning [Lan14].

Conclusion

In the scope of this dissertation, no additional sources regarding the theoretical foundations of game-based learning, serious games, and gamification are considered. The focus of this work lies less on the theoretical foundations and more on applied research towards improving game design and the application of personalization.

2.2.4 Learning Games

Definitions

The term *learning game* denotes a subgroup of serious games, focusing on informal learning [Dör+16], while the term *educational game* is used for serious games in formal educational contexts (e.g., university). In the scope of this dissertation, the focus is on learning games for anti-phishing education. Although anti-phishing education could be part of a security-related curriculum in formal educational contexts, it also addresses the general public. As such, anti-phishing learning games could be used in both formal and informal educational contexts, making this distinction between educational games and learning games less precise. To this end, it can be observed that both terms are often used interchangeably in literature.

Learning game design

As learning games are used for educational purposes, the design of learning games is a balancing act between the gaming and the learning aspect [KKS11]. While the game design may involve a notion of interactivity and immersion, the design of learning interventions may involve some kind of educational framework for selecting learning content and the formulation of learning outcomes or goals. Thus, the design of learning games could utilize both approaches from game design and educational frameworks. In [KKS11], Kelle et al. explain how game design patterns can be linked with educational functions to give a perspective on pedagogically founded and well-structured learning game design. While [KKS11] presents a set of game design patterns and maps them to different aspects in educational design processes, learning game design in the scope of this dissertation focuses more on the definition of learning goals and the selection of learning content.

Learning game development

For the development of learning games, different frameworks, and technologies can be used. Similar to commercial games, learning games can be implemented for various platforms and using different game development frameworks. One prominent example is Unity[1], a real-time 3D development platform for building 2D and 3D applications. It can target various platforms like mobile, desktop, different consoles, VR, and the web. Depending on the type of game, the underlying platform, and the desired functionalities, different frameworks are more or less suitable. When focusing on games for the web, one particular framework is the Multi-Touch Learning Game framework (MTLG), developed and maintained as an open-source project[2] by the Learning Technologies Research Group at RWTH Aachen University. While this framework was initially developed for

[1] https://unity.com/, accessed on 22.12.2021
[2] https://mtlg-framework.gitlab.io/, accessed on 22.12.2021

the creation of collaborative games for large-scale multi-touch tabletop displays, the framework utilizes the HTML Canvas element and modern JavaScript (JS) [ERS18]. The framework is based on CreateJS[3], a suite of modular JS libraries to enable rich interactive content via HTML. Thus, it can be used to develop games for any browser-capable device as well as any screen resolution. Compared to Unity, the MTLG framework provides various features to support learning game development (e.g., a learning analytics logging module, a feedback module [Küh18], and a remote-teacher module [3]). For the scope of this work, the MTLG framework is used as it supports the development of web-based learning games. Furthermore, future developments of the framework can be easily integrated in the research process of this dissertation as they may benefit the game development.

Game Learning Analytics

In order to gain insights on how playing a learning game may support the learning process, data-driven approaches like game learning analytics can be used. *Game learning analytics* (GLA) combines the educational goals of learning analytics with the methods and tools from game analytics [Fre+16]. The objective is to provide insights into the gameplay sessions and provide detailed interactions about how players interact with the game [Fre+16]. In the context of this dissertation, *learning analytics* (LA) is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" [Sie10] and *game analytics* (GA) describe the application of analytics to game development or game research to understand better how players behave in the games as well as to find errors and improve the game [SDC13; Fre+16]. As such, game learning analytics can be defined as the measurement, collection, analysis, and reporting of gameplay data to understand and optimize the learning process and experience, which includes the learning game itself.

2.3 Personalization

A central research area this dissertation touches is personalization in learning and closely related domains. As such, this section presents a definition of personalization and related terms. Further, objectives, requirements, and methods for personalization are briefly introduced to allow this work to build upon it.

2.3.1 Definitions

The idea of personalization is rather old and has been proposed in various contexts by different stakeholders. It shares different definitions for different communities, and as described by Fan and Poole, the concept can be "intuitive but also slippery" [FP06, p. 183].

Origin

In 1956, Smith introduced the concept of *market segmentation* as a solution to the problem of heterogeneity of the market. It describes the adjustment of products and marketing strategies according to the needs and requirements of consumers by viewing the market as smaller, more homogeneous markets [Smi56]. Although not called personalization, market segmentation is a similar concept in the context of commerce, as it describes tailoring marketing strategies to the needs of smaller consumer groups. Also, Churchill describes personalization as an extension of market segmentation since in the

[3] `https://createjs.com/`, accessed on 07.01.2022

field of Human Computer Interaction (HCI), "personalization is largely about filtering content to satisfy an individual's particular tastes" [Chu13, p. 12].

Key aspects in definitions

Strongly driven by technological innovation, personalization can be applied to various domains, and increasing interest is shown by academia and industry. Fan and Poole compared different definitions of personalization and concluded that most definitions include the following three key aspects: (1) the purpose of personalization, (2) what is personalized, and (3) the target of personalization [FP06]. As a result, their proposed definition specifies "personalization as a process that changes the functionality, interface, information access, and content, or distinctiveness of a system to increase its relevance to an individual or a category of individuals" [FP06, p. 183].

Clarification of related terms

In [SS16], definitions for personalization and other related terms, such as customization, adaptability, and adaptivity, are presented in the context of serious games. It presents a foundation for this dissertation as it defines essential terms of the respective research area. Meanwhile, in other literature, the terms personalization and adaptivity are often used interchangeably, so it is even more important to clarify the understanding in the scope of this work. According to [SS16], personalization and customization are related terms but differ in their understanding of the target of personalization. Personalization is defined as "the act of changing a system to the needs of a specific individual user" [SS16, p. 344]. While personalization focuses on the needs of specific individuals, the concept of customization is defined "the act of changing a system to the needs of a user group" [SS16, p. 344]. This difference is subtle, but customization is coarser since it focuses on user groups instead of individuals. Besides, both concepts require adaptability, which describes that systems are not fixed but can be changed [SS16]. Lastly, adaptivity is defined as the fact that a system dynamically changes over time, and thus, it is not fixed [SS16].

Application in games

Often personalization and adaptivity are used in similar contexts, but [SS16] presents a clear distinction between both concepts. While personalization utilizes change to accommodate an individual, adaptivity describes the change that happens automatically. When applied to serious games, a personalized game is adapted to the player's characteristics and needs [SS16]. This adaptation can be made manually or automatically. However, it may differ from an adaptive game, which includes automatic adjustments of the game over time [SS16].

Conclusively, the work in this dissertation follows the understanding of personalization presented in [SS16] and defines it in the games context as the act of changing a system, i.e., the learning game, to fit the requirements and needs of individual learners.

2.3.2 Objectives

While personalization usually targets the users of a system, the reasons for personalization are often domain-specific. Besides targeting the system's users, it can also benefit the system's designers, maintainers, hosts, or even third parties. This depends on the overall objectives of why to implement personalization.

Optimization

The most common objective of personalization is the optimization of a system towards domain-specific goals. Especially in application areas like online retailing, search engines, and recommender systems, optimization is a crucial factor to success. Whether it is the personalization of advertisements, search results, or product recommendations, personalization has the objective to optimize towards specified goals and offers "a potential source of competitive advantage" [Kar+17, p. 370].

Paradigm shift

Another objective derives from a paradigm shift: While systems were implemented as a one-fits-all solution in the past, today, the perspective differs. Through personalization and customization, systems can be tailored to their users, and this may result in increased usage and increased user acceptance [SS16]. [OOT17] presents a comparative study of system-controlled and user-controlled personalization approaches and identified seven common strengths. These strengths can be used to derive further personalization objectives, such as increasing the usefulness, ease of use, and relevance. If the user is explicitly involved in the personalization process (user-controlled personalization), it supports self-efficacy, trust, the feeling of control, and freedom [OOT17].

Game-related objectives

In the domain of games, personalization yields the potential for increased involvement and increased player satisfaction [BTP12]. Since personalization is implemented in more and more commercial video games, players may often expect personalized gaming experiences. In [BTP12], a discussion on the motivations and objectives of personalization in games is presented by drawing arguments from psychological research and the domains of game design and game development. It is argued that a fit between the personality and the environment can raise player loyalty and enjoyment [BTP12].

Learning-related objectives

The before-mentioned paradigm shift also applies to the domain of education, where it shifts from traditional one-fits-all teaching approaches to adaptive and personalized learning. Personalization and adaptation can support heterogeneous learners by tailoring the learning environment with its content, methods, and objectives. The objectives include maximizing satisfaction, learning speed, and learning effectiveness. [SS16] argues that personalized learning games could support learners in making progress in a motivating and rewarding way to reach desired learning outcomes. Personalized learning games could hence increase motivation and offer a personalized user experience.

Overall, the objectives of personalization vary between different domains and application areas. It can offer a competitive advantage and influence the relationship between users and a system. While personalization allows for more individual support in the learning domain, it is used to increase player loyalty in games. Within this dissertation, personalization offers a potential solution to increase the relevance of learning game content. The objectives are to provide a personalized user experience and increase usefulness and user acceptance. Additionally, an increase in motivation and satisfaction may also be objectives.

2.3.3 Requirements

As mentioned in Chapter 2.3.1 personalization requires adaptability. A system needs to be adaptable to support personalization [SS16]. On the contrary, personalization cannot be implemented when a system is fixed and does not allow any kind of adaptation. This basic requirement allows for personalization in various domains. However, as this dissertation focuses on personalization in learning, domain-specific requirements need to be discussed.

Learning-related requirements

Next, personalization in learning primarily requires two sets of inputs: learning resources and learner data [IB18]. While in personalized online courses, the learning resources are different types of resources (e.g., videos, slides, documents) usually provided by an instructor, in learning games, it heavily depends on the type and topic of the game (e.g., images, videos). Learner data is all data characterizing a learner. This can include data on a learner's interests, personality traits, learning style, past learning activities, and achieved results in assessments and interactions within a learning environment. Within learning games, learner data may also include gameplay-related

data such as scores, in-game decisions, and detailed activity logs. The granularity of learner data can vary depending on the system. While learning analytics components in modern learning management systems capture more and more learning experiences via log data, they are often bound to page visits and click events. In games, logging capabilities allow more detailed tracking of a player's actions.

Architectural requirements

Diving deeper into architectural aspects of personalized learning systems, [IB18] identifies four primary functions resulting in four main components of a personalized learning system:

- learner unit,
- knowledge unit,
- personalization unit, and
- presentation unit.

Each component may have different implementations depending on the context and fulfills a required functionality for personalization [IB18]. While the learner unit captures data about the learner, the knowledge unit maintains learning resources within the system. The personalization unit is responsible for learner modeling and mapping learning resources to learner models [IB18]. In this context, a learner model refers to an abstract, structured representation of the learner's knowledge, misconceptions, and difficulties [Bul04]. Lastly, the presentation unit is the environment in which learning resources are delivered to the learner (e.g., the learning game) [IB18].

Overall, the requirements for personalization are threefold: First, personalization requires the adaptability of a system. Second, personalization in learning requires learner data and learning resources. As such, a mapping of learning resources to individual learners can be implemented using a learner model. Third, additional requirements can be considered depending on the application area within the learning domain.

2.3.4 Methods

The process of personalization has been outlined and described for various domains, but it is usually divided into three phases: (1) Learning, (2) Matching, and (3) Recommendation [Chu13]. For each phase, various methods fulfilling domain-specific requirements exist.

Methods for Learning

Starting with the 'learning' phase, it can be distinguished between implicit and explicit methods. For *implicit learning*, automated, non-invasive techniques are used to collect information about the user, i.e., the learner, in the context of this dissertation. Depending on the system, this may include various data sources, e.g., activity logs or user profiles. *Explicit learning* usually involves active participation by the user, i.e., by providing the relevant information through user input. This can be done using questionnaires or user-generated signals, e.g., ratings or 'likes' used in social media platforms. [Chu13]

Methods for Matching

For the 'matching' phase, often domain-specific methods for matching are used. While in recommender systems, methods, like collaborative or content-based filtering, are applied in a hybrid fashion, traditional intelligent tutoring and expert systems rely on rule-based filtering to provide recommendations to their users [Chu13]. In more modern systems, the 'matching' might be done using AI technology and Machine Learning algorithms. Overall, the most contemporary personalized systems implement hybrid approaches combining different methods for matching.

Methods for Recommendation

Lastly, in the 'recommendation' phase, the results of the previous matching need to be presented to the user. Here, the delivery or recommendation methods vary depending on the application area (e.g., in an online shop, product recommendations are presented at various points during a user's visit). In game-based learning, recommendations could be delivered as personalized game content or instructions.

Process and outcome personalization

Another distinction are the interrelated concepts of outcome personalization and process personalization [Chu13]. The meaning of both can be easily explained using examples: *Outcome personalization* is operationalized when a user who searches for cameras in an online shop gets recommended other photography-related products. Likewise, when visiting theme parks or chain restaurants, *process personalization* can be encountered through the same greetings, the same uniforms, the same interaction patterns.

While the personalization target is usually a system's user, there are different personalization types or aspects that can be adapted, e.g., the language of a user interface. In the broad context of information systems, Fan and Poole distinguish between four different types of personalization [FP06]:

- content
- user interface
- channel/information access, and
- functionality.

Since learning games can be understood as a subtype of information systems, the distinction of [FP06] can be applied to the context of learning games without complex adaptations. The personalization of content and the user interface is rather intuitive, e.g., through adaptation of the tasks within the learning game (personalization of content) or the language of the game's user interface (personalization of the user interface). When personalization is applied to the channel or information access, it is an adaptation of the media type used to deliver information, e.g., animation sequences or digital storytelling within a game. Lastly, the personalization of functionality impacts a player's actions and controls.

Conclusively, in the context of this dissertation, the three phases described in [Chu13] need to be applied to the domain of learning games for security education. In addition, the utilization of various methods and different personalization types needs to be considered for the design and implementation of new, custom methods for the respective domain.

3 Related Work

This chapter presents a review of related work in the field of personalized anti-phishing learning games. Prior to this work and its previously published papers, the research domain of personalized anti-phishing learning games has not been systematically explored, i.e., no researchers have published about possible approaches to personalize anti-phishing learning games. In the broad field of phishing research, researchers might have explored personalization for technical countermeasures and detection systems. However, this does not lie in the focus of this work.

Review of neighboring research fields

Since no immediate related work in the narrow field exists, neighboring research fields are reviewed to establish the state of the art and provide insights into recent results and developments. This includes mainly three dimensions to look into: (1) personalization in security education, (2) personalization in learning games, and (3) personalization in learning in general. Lastly, a summary and conclusion on the state of the art are presented to provide value to this work and the developments described in later chapters.

It should be noted that research about game-based anti-phishing education and game-based security education is not further discussed in this chapter. Instead, Chapter 5 presents two systematic literature reviews and a deeper analysis of existing games and prior work in these domains.

3.1 Personalization in Security Education

In this section, a brief overview of existing work in the field of security education is presented. The focus lies on personalization in security education. This overview aims to provide insights into the current state of the art and present findings relevant to this dissertation.

Over the last two decades, various researchers and organizations have explored different approaches to security education. Similarly, different review publications about security education have been presented [AS18; ABS17; Qua21]. These reviews often provide an overview of the challenges and trends in security education. However, they can also be focused on the methods and topics in security education. Thus, possible work on personalization in security education might be reflected in those publications.

Reviews on security education

While there exist various reviews, even for particular topics (e.g., social engineering [AS18], phishing [ABS17]) or target groups (e.g., children [Qua21]) in security education, there are no mentions of 'personalization' or related concepts. This indicates that personalized security education has not been extensively researched and reviewed by other researchers when establishing the state of the art of the research domain. However, a look into further literature beyond the existing reviews yields some related work in the field of personalization of security education [VF18; VF19; Tal14].

Conceptual work

In [VF18], it is proposed to tailor security education towards individual learners instead of delivering one-fits-all solutions. While one-fits-all solutions at least serve to ensure that learners have the opportunity to learn about security, it does not ensure

proper awareness and understanding of security. For the personalization of security education, [VF18] suggests utilizing the threshold concepts and troublesome knowledge by [ML03] to help educators to understand learning barriers and reconsider their way of delivering security education. Lastly, [VF18] introduces a framework for individual security learning, in which different aspects (e.g., learning styles, prior knowledge, learning barriers) are considered when creating and delivering security education. This way, security education moves from a one-fits-all approach to a tailored, personalized approach.

Although [VF18] considers various aspects of personalizing security education, the work is only a proposal. It does not present any implementation or example of tailored security education [VF18]. Also, the scope of the work is organizational security education within a company where security managers responsible for a SETA program should consider tailoring their offers to employees. This way, it could deliver a more tailored security experience and meet the employees' needs. [VF18] suggests upfront data gathering using a questionnaire as a diagnostic tool to determine where their employees are currently at and to set individual starting points for employees. Unfortunately, neither a questionnaire nor any details about how to collect valuable data are presented in [VF18].

In a follow-up publication, the framework is extended by a peer-learning component. [VF19] identifies the significant need for a community of good practice for security education. Therefore, it is proposed to form peer-based learning relationships between employees and proactive steering by designated peer-based leaders who are more engaged in promoting security education. The benefits of peer learning for security education may include promoting continuous learning among colleagues and a cost-effective approach to post-training phases (i.e., if after initial training, employees keep learning from each other in behaving risk-aware and securely). However, again, the follow-up work does only provide a proposal of peer learning and no actual implementation [VF19].

PISE framework

Besides the described conceptual work, the *Personalising Information Security Education* framework (PISE) presents a holistic approach to personalize security education based one individual's learning style [Tal14]. PISE is implemented as a web-based system and is designed for two contexts: (1) organizational contexts and (2) the general public. The system offers different learning style models (e.g., the VARK - Visual, Aural, Read/write and Kinaesthetic - model [FM92], the Felder-Silverman Learning style model [FS88]) and provides a complete learning environment including learning materials and subsequent assessments. Beyond learning styles, PISE can utilize further personal information and prior knowledge to deliver a personalized learning plan. The personalized learning plan can also be generated using a given syllabus if used in organizational contexts.

In [Tal14], the design and implementation of PISE are described in detail; however, evaluation of the system does not go beyond a qualitative review by a domain expert. User evaluation with either the employees in an organizational context or private end-users in an informal learning context is still needed to gain insights on how personalized security education based on learning styles affects secure and risk-aware behavior of users. As the concept of learning styles is often criticized [Kir17], it is unclear whether PISE provides a suitable solution to personalizing security education and whether possible effects can be attributed to different learning styles [Tal14].

Further work using learning styles

Besides [Tal14], [AAH15] also applies the concept of learning styles to security education by providing a personalized sequence of learning materials based on the Felder-Silverman Learning Style model. A study with 60 participants revealed positive effects

on learning gain and student satisfaction. [AAH15] concludes that the approach is beneficial to security and could also yield benefits in other CS-related topics. However, the study's limitations include the small sample size and a limited number of learning materials [AAH15]. Thus, it limits the power of personalization, i.e., the variations between different sequences are limited to the number of materials. Similar to [Tal14], learning styles might not be the reason behind these findings, and thus, the results should be discussed carefully. However, as this dissertation does not utilize learning styles for personalization, it will not be further discussed.

Compared to general learning style models and their respective inventories for measuring, [Pat+18] modifies the inventory of the VARK model and created the 'Cyber-security Learning Styles Inventory'. While the inventory still measures the learning styles according to the VARK model [FM92], it incorporates security-related items to provide a more suitable context. As such, [Pat+18] presents an online study using the modified inventory. The study's objective was to investigate whether security education adapted to individual learning styles would lead to a higher level of awareness. The study revealed a mismatch between the modes of training the participants received compared to the modes they would have preferred according to their learning style. Furthermore, the study showed that a participant's learning style was positively associated with the level of awareness and thus, the authors' hypothesis was supported [Pat+18]. While this study was only based on a survey and no implementation of personalized security education was evaluated, the actual impact of how personalization can affect the level of awareness is not yet determined.

Adaptivity in security education

Since personalization is often implemented using adaptivity, it is worth considering prior work in the field of adaptive security education. Therefore, adaptive educational interventions (e.g., learning games) in the field of security education were reviewed.

[TT20] presents a study of player behavior and electroencephalography (EEG) headset readings during the gameplay of *Brute Force*, a learning game on password security. The game implements dynamic difficulty adjustment (DDA), in which the game's difficulty steadily increases until players lose. Each difficulty adjustment is computed using a challenge function. The objective is to foster a continual state of flow as defined by Csíkszentmihályi [Csi97]. [TT20] evaluates the players' behavior in the game and correlated it with the stress level measured using EEG readings. While the work in [TT20] is somewhat exploratory, future work opportunities include the integration of players' EEG readings as input to the DDA implementation.

Another example of adaptivity in security education is presented in [SLJ17]. In university or school contexts, security education often integrates different exercises and tasks for students to apply theoretical knowledge and practice other skills. Security challenges similar to those offered in CTFs competitions and other events can be used as exercises for students. However, since these security challenges can vary in difficulty levels, it can be problematic for students to select appropriate challenges based on their skill level. Therefore, [SLJ17] proposes an adaptive approach using a rating system to rank the challenges and the students and provide guidance in choosing suitable challenges. Compared to DDA, this approach is more manual and provides adaptivity through user input, i.e., choosing a suitably ranked challenge. Nevertheless, it is a valid approach to personalize security education. Since the work in [SLJ17] is only theoretical, it still needs to be implemented and evaluated to gain insights on how rating systems can be beneficial to security education.

Conclusion

In conclusion, the reviewed work in the field of personalized security education focuses more on conceptual approaches [VF18; VF19] and less on implemented, evaluated work.

To this end, mostly learning style models were considered for personalization [AAH15; Pat+18; Tal14]. Regarding adaptivity as an approach to achieve personalization, rating systems matching participants' skills and the challenge according to difficulty [SLJ17] as well as the adaptation of gameplay to foster player engagement and a state of flow [TT20] can be found in related work. For personalization on a content level, no related work focuses on personalized content in security education. Only [SLJ17] could be considered since it matches the learners' skill level to the difficulty of challenges. However, since it is only conceptual work, no implementation or existing system could be used to extend and adapt to the work of this dissertation.

3.2 Personalization in Learning Games

While personalized learning games in the domain of security education have not yet been widely explored, researchers have studied different approaches for personalization in learning games in other domains. This section presents an overview of existing work, its achievements, and implications for future research. Similar to the previous section, different examples of personalization and adaptivity in learning games are reviewed. As noted before, the terms personalization and adaptivity are often used interchangeably (see Chapter 2.3.1).

Key concepts and approaches

An appropriate starting point when reviewing the literature on personalized learning games is the work presented in [SS16] and [GW16]. Besides introducing definitions for key terms like personalization and adaptivity, [SS16] presents the state of the art at the time and review various publications of fellow researchers. Furthermore, neighboring domains like Artificial Intelligence (AI) in games and Intelligent Tutoring Systems (ITS) are discussed, as they are often connected to either personalization or games, and research domains may overlap. Focusing on learning games, [SS16] emphasizes the research on DDA in games as well as dynamic adjustment of game mechanics or content (e.g., personalized scenarios in games [NR09]). [GW16] presents definitions for related concepts and describes models and mechanisms for the development of personalized, adaptive learning games. For adaptation in learning games, three dimensions are described: (1) difficulty adaptation, (2) player modeling, and (3) learner modeling. In addition, different concepts and algorithms used in learning games are presented as state of the art. The main contribution in [GW16] is, however, the concepts of adaptive storytelling and narrative game-based learning objects (NGLOBs). NGLOBs represent small, atomic units of contextual information in learning games and encode a composition of narrative, gaming, and learning objects. Furthermore, [GW16] introduces the research project *80days*, which uses NGLOBs to provide adaptive storytelling in a learning game about geography. In conclusion, the work presented in [SS16] and [GW16], both provides insights into the research domain by establishing a common understanding of relevant terms and concepts and presenting approaches and methods used in related work and the state of the art.

As learning games are a sub-genre of games, research in the field of personalized gaming may also enrich the state of the art. [BTP12] presents an overview of personalized gaming, its motivation, and possible components of a personalized game. They distinguish between the following eight different adaptive components:

- Space adaptation
- Mission/task adaptation
- Character adaptation

- Game mechanics adaptation
- Narrative adaptation
- Music/sound adaptation
- Player matching
- Difficulty scaling

Learner models

Depending on the game, each of these components could also be utilized to implement personalization in learning games. The base for personalization and input to these adaptation components is the learner model, which is a representation of the learner providing specific information that can be used to steer personalization, e.g., a measured skill level that serves as input for difficulty scaling within a game. In literature, different parameters or models are included in the learner model and used for personalization, e.g. gender [SKA09], personality traits [Tli+19], learning styles [Hwa+12], emotional states [BV16], spatial visualization skills [Xia+18] or in-game performance [Pla+19].

Project ELEKTRA

Among research projects on adaptive and personalized learning games, the *ELEKTRA* project was an interdisciplinary project of the European research community, in which a personalized 3D adventure game for teaching physics according to national curricula was developed using a sound psychological and pedagogical framework [Kic+06]. The game aims to teach optics, a topic in physics curricula and is designed for students ages 12 to 13 (i.e., eighth grade). Within the game, personalization is achieved using adaptive gameplay based on the learners' prior and current knowledge. In particular, micro-adaptivity is implemented, i.e., "adaptation within learning situations as opposed to around them" [Kic+07b, p. 651]. The authors argue that continuous and less periodic adaptation at a higher frequency than between tasks is needed to support immersive 3D games. However, the implementation of micro-adaptivity needs to be in consideration of the gaming experience. Therefore, the *ELEKTRA* project uses a recommendation system that provides recommended changes that are contextually specific and in line with the current gaming experience. The final realization of a recommendation is lastly determined by the game while ensuring that it does not compromise the gaming experience. In [Kic+07a], the underlying concepts and models used in the *ELEKTRA* project are described, and the architecture for micro-adaptive assessment and interventions is presented. In further publications, more details about the implementation of micro-adaptivity are presented using a specific ontology model [KA08] as well as the actual game prototype [Kic+08].

Project 80days

A successor research project is the *80days* project [KGA08]. Similar to the *ELEKTRA* project, it was an interdisciplinary project funded by the European Commission and ran from 2008 to 2010. The game's theme was inspired by Jules Verne's novel 'Around the world in eighty days' and it aims to teach geography to a target audience of 12 to 14-year-old students. While the previous project implemented micro-adaptivity, in the *80days* project, macro-adaptivity was implemented. Macro-adaptations are based on a fixer learner or adaptation model and more traditional adaptation techniques are used (e.g., adaptive presentation or navigation) [KGA08; Kic+11]. In *80days*, macro-adaptivity is implemented using adaptive storytelling through game paths [KGA08].

Micro- and macro-adaptivity

While the concepts of micro-adaptivity and macro-adaptivity can be considered for the games of this dissertation, the overall focus of both the *ELEKTRA* project and the *80days* project was more on immersive learning games using a 3D game environment and digital storytelling. However, immersion is not a key aspect in this dissertation, and no 3D game environment is used. Lastly, micro- or macro-adaptivity was not utilized in this work since personalization was not implemented through adaptivity during gameplay.

Nevertheless, when introducing adaptivity to anti-phishing learning games, it should be implemented in consideration of the gaming experience, and therefore, the concepts of micro- and macro-adaptivity should be revisited.

Content personalization

Beyond the concepts of micro- and macro-adaptivity, research on personalized games also focuses on different types of games, e.g., personalized exergames, a subgenre of serious games for sports and health. [Göb+10] introduces a set of personalized exergames which combine concepts of adaptation, personalization, and authoring. First, doctors, trainers, or fitness coaches can create personalized training plans through an authoring component. Next, an adaptive engine implements these personal training plans via adapting the gameplay. Different sensor data like vital status, pedal resistance, and movement data are used as input for the learner model and, thus, the adaptation of the game. While the adaptations based on sensor data fall into the category of adaptivity, the use of an authoring tool to create personalized training plans is actual content personalization. This way, the gameplay differs between players depending on the exercises a trainer selects with the authoring tool. Beyond [Göb+10], further research in the health domain focus more on DDA and less on actual content personalization (e.g., [Hoc19], [MDZ14]). Theoretical work in the field of pervasive health games explores personalization based on gamer types [Orj+13].

Content generation

In the field of content personalization in games, the question of content generation arises. Here, different types of content, as well as methods for content generation, can be considered [DK16]. Since content creation is a labor and time-intensive task for game developers, automation of content generation is a desirable goal. Content, i.e., "any piece of data that a game loads and uses in the process of typical gameplay" [DK16, p. 174], can be divided by its type. This includes media for visual or aural aesthetics or content for specific game mechanics (e.g., pieces for a puzzle game). Furthermore, story elements like quests or tasks, including objectives and conditions, are also content. For content generation, [DK16] suggests the distinction of developer content (DC), procedural content (PC), and user-generated content (UGC). These types are described as follows [DK16]:

- DC includes the initial, tailor-made content of developers, which is usually of very high quality as developers know about the game's structure and design and the potential limitations and constraints of the underlying game engine or framework. The standards for this type of content are very high, as this content is used to 'hook' players. DC can be considered time and labor-intensive, so automation is desirable.
- PC describes algorithmically generated content using different automated or semi-automated methods. So-called procedural content generation (PCG) allows for the content creation "that can build upon a baseline and produce a very large number of variations" [DK16, p. 177]. Methods include noise types like attribute randomization to grammar-based generation with constraint satisfaction. While automation can save time and effort, fully automated PCG can create content that may differ from the developers' original vision. Here, procedural assistance can provide a balance between automation and user control. Even some methods of AI have been used for PCG[4].
- As time and budgets for learning game development are limited, UGC can also be considered. While this concept has some examples within the commercial gaming sector, there is also a potential for serious games, particularly the educational

[4] As noted in [DK16], PCG has even reached a point where even the generators themselves can be generated procedurally.

domain. UGC would allow for creative processes where users contribute a created artifact or asset (due to a game task). Furthermore, it could contribute to specializations and variations of the game through new levels or even personalization (e.g., personal avatars, personalized game worlds).

Beyond this distinction, [DK16] also presents different examples for content generation as part of the state of the art. It presents examples for scenario or terrain generation that are less relevant in the scope of this dissertation but also emphasize that procedural generation can be taken a step further when creating content that dynamically adapts to the user. For this dissertation and the goal of personalizing anti-phishing learning games, PC and UGC are interesting concepts that can be considered during designing and implementing a personalization framework. Especially, PCG would allow for automated generation of different game content (e.g., phishing URLs or emails), and when based on the learner model, generated content could even be personalized. Here, DC would be limited and less flexible since developers need to anticipate different players to provide personalized content. Therefore, DC could best be utilized for customization, i.e., the adaptation of game content based on user groups.

Conclusion

In conclusion, existing work in the field of personalized learning games is strongly driven by adaptivity, as games are dynamic environments that allow incremental adaptation during gameplay. The projects *ELEKTRA* and *80days* systematically explore different types of adaptivity, and present the variety behind personalization, since it can be done within by adapting gameplay within levels or on a high level through storytelling and navigation. Outside the educational domain, personalized gaming utilizes different components to implement adaptivity. Depending on the game, they can be utilized for learning games as well. Examples beyond adaptivity include content personalization of exergames, where personalized training plans are provided for individual players. Since no prior work on personalized anti-phishing learning games exists, related work is drawn from the neighboring domains. As such, the work presented in [SS16] and [GW16] are both considered for this work. Both publications establish definitions for personalization and adaptivity and present significant related work of the research domain. Furthermore, they present different approaches and methods and discuss the benefits and challenges of personalization in learning games. For the personalization of content, related work in the field of procedural content generation can be considered as it yields the power of automation and, paired with suitable learner models, the generation of personalized game content (e.g., [DK16]). Thus, suitable methods of PCG are considered during the design and implementation of the personalization framework in this dissertation.

3.3 Further Directions in Personalized Learning

In this section, the review of related work is completed by briefly assessing the state of the art of personalized learning and its further directions beyond learning games or the application area of security education. As such, this section will only present an overview provided by review papers (e.g., [LW21], [LW19], [Xie+19]).

Personalization in education

Throughout the years, various researchers presented their ideas, approaches, and findings in the field of personalized learning. As such, [Cam+07] depicts the origins of personalization and transferred them to the educational sectors. [Cam+07] summarizes the possible benefits of enabling students to set their own goals and assess their own progress and learning flexibility in terms of time and place while promoting self-regulation and motivation. Here, the actual benefits may differ depending on the

specific learning environment in which personalization is implemented. In the context of mobile learning, [Ber11] reviews the potentials of personalized learning since it provides choices in what, how, and where to learn. While these two publications are more than a decade old, the advancements in technology until today call for a more updated review of personalized learning.

Reviews on personalized learning

Furthermore, two comprehensive reviews on the development of personalized learning over time, and future research directions are presented in [LW19] and [LW21]. While [LW19] reviews the literature from ten years between 2009 to 2018 and summarizes how learning had been personalized so far, the second review extends the reviewed time span (between 2001 and 2018) and focuses solely on journal publications [LW21]. The most common research issues are the effectiveness of personalized learning and learners' satisfaction [LW19; LW21]. Within the last decade, research regarding the acceptance of personalized learning and learner engagement in personalized learning context increased. Among the means to achieve personalized learning, Li and Wong found that more than a third of reviewed publications describe the development of intelligent learning systems, which include new technologies (e.g., AI, semantic web) and consider learners' individual characteristics (e.g., personality traits, prior knowledge). Other means include intelligent tutoring systems, learning management systems, mobile learning, or flexible curricula (i.e., curricula where the course content and delivery is flexibly customized based on learners' individual needs and progress). It should be noted that since learning management systems nowadays allow for customization and extension through connected services or technologies (e.g., Moodle with its large plugin directory[5]), it can also be considered an intelligent learning system. Similarly, this applies to intelligent tutoring systems, which can be integrated into learning management systems or part of an intelligent learning system. In confirmation of the findings presented in the previous section, game-based learning was also identified as a means to achieve personalized learning. Furthermore, the reviews analyzed the most frequent aspects which were personalized in existing studies [LW19; LW21]. Here, learning methods account for more than 50 percent of the reviewed literature, followed by personalized materials. Regarding the objectives of personalized learning, the reviews identified that most studies focused on the increase of learning effectiveness as well as providing personalized learning paths and caring for learners' motivation, engagement, satisfaction, and an enriched learning experience [LW19; LW21].

Besides the extensive review work presented in [LW19] and [LW21], [Xie+19] also presents a review of trends and developments in personalized learning while focusing on journal publications between 2007 and 2017. While this work might be similar and somewhat redundant with [LW21], it identifies different learning support types for personalization (i.e., the component which implements personalization in a specific context) and also provides a discussion of potential research issues and trends for future work. As such, [Xie+19] identifies that learning content is personalized in most studies, followed by personalized learning paths. Furthermore, some studies implemented personalized interfaces, suggestions, prompts, and feedback messages. Regarding the research trends for personalized learning, [Xie+19] suggests exploring other domains by personalizing existing domain-specific learning systems, e.g., using knowledge graphs to gain domain knowledge and automatically (re-)structure learning materials. While most studies implemented personalized learning on traditional computers, the advent of mobile and wearable devices (e.g., smartwatches) allows new personalized learning approaches, which can utilize different sensor data and adapt the learning environment accessed through different devices. Lastly, with the increase of collaborative and

[5] https://moodle.org/plugins/, accessed on 21.09.2021

immersive learning environments (e.g., VR), personalized learning can go beyond the selection and presentation of learning content. Studies can investigate the potential of personalized learning for higher-order thinking skills and communication skills.

For the scope of this dissertation, the focus lies on personalized learning in game-based learning and the domain of anti-phishing education. The design and implementation of a personalization framework are tailored to the game environment; however, possible transfers of framework components should not be dismissed. Similar to common trends within the last two decades, learning content will be personalized as part of this work. While DDA could also be explored, personalization does not need to be implemented using adaptivity.

3.4 Conclusion and Identification of Research Gaps

This section presents a brief conclusion of reviewed related work and identifies potential research gaps. Related work will guide the design and implementation of learning game prototypes and the development of a personalization framework for anti-phishing learning games. As initially stated, a review of existing research in the domain of anti-phishing learning games did not yield any prior work regarding personalization. The research domain has not yet been systematically explored, which is why no immediate related work was available for review. Thus, the review focused on related work in the neighboring research fields.

Personalization in security education

For security education, different personalization approaches have been suggested as part of theoretical or conceptual work or implemented in different learning contexts. Only a few implementations have been found in literature, e.g., the PISE framework, a personalized learning framework that uses learning styles to provide personalized learning plans with suitable materials and subsequent assessments [Tal14]. Although heavily criticized nowadays, the concept of learning styles was also used in other studies [AAH15; Pat+18]. Beyond the use of learning styles, personalized security education was also implemented through adaptivity, e.g., DDA in *Brute Force*, a learning game about password security [TT20]. Another example uses an adaptive approach for security exercises in a course context [SLJ17]. Here, a rating system matches difficulty with skill level and provides suitable exercises to students. However, the personalization through learning styles or DDA does not address the personalization of learning content by content generation, but rather the access to existing learning content. In addition, conceptual or theoretical work only provides the idea and potential approaches to realizing it. Still, it does not yield any implementation to use and extend for personalizing anti-phishing learning games. Conclusively, related work in the field of personalized security education yields proof that it has been done before and has even been studied using an adaptive learning game for password security. However, there is no prior work on how to personalize learning content for anti-phishing education. This presents a research gap, not only in the particular context of anti-phishing learning games but also in security education.

Personalization in learning games

In the domain of personalized learning games, related work includes contributions regarding key terms and concepts and representative projects that implement personalization through different means. While in [SS16] and [GW16], foundations for the conceptual understanding of personalization, adaptation and adaptivity in the context of learning games is presented, the projects *ELEKTRA* or *80days* present the actual implementation of personalized learning game prototypes [Kic+06; KGA08]. Although work on clarifying key terms exists, the review of related work confirms that personalization

and adaptivity are often used interchangeably. As such, concepts like micro-adaptivity and macro-adaptivity are implemented to personalize learning games. Compared to the personalization of learning content, both approaches are more about personalized access to learning content than their generation. However, content personalization can be found in the subgenre of exergames, e.g., in [Göb+10] an authoring tool is used to provide personalized training plans as input to a set of games. Furthermore, more related work about content generation can be found when the scope is extended beyond learning games or serious games. In [BTP12], personalized gaming is described, and different adaptive components are presented, which can be integrated into other learning games as well. Regarding content generation, the distinction between developer, procedural and user-generated content is important to consider (as presented in [DK16]). PCG can be used for personalized learning games, where algorithmic power dynamically generates content based on a learner model. Finally, this reveals a research gap since no PCG has yet been explored for the personalization of anti-phishing learning games. Whether it is the design of a PCG-based personalization framework or the implementation and subsequent study of the effectiveness of procedurally generated content in personalized learning games, it could provide meaningful insights and findings in the research domain of personalized anti-phishing learning games.

As stated before, this chapter does not elaborate on research about game-based anti-phishing education, although it is a connected research domain to this dissertation. However, in Chapter 5, two systematic literature reviews provide a more in-depth study of related work and existing games. Therefore, the identification of research gaps will be extended by this chapter to provide input to the design and implementation of own learning game prototypes.

4 Research Design and Methods

This chapter presents the research design this work follows and serves as a methodological overview to understand the following chapters. The subsequent sections outline the different steps of the research process as well as methods and tools utilized throughout this work.

4.1 Methodological Approach

The methodological approach in this dissertation is applied research with a combination of multiple types of research designs. It follows a combination of design-based research and experimental hypothesis-testing research using different methods and tools to fulfill the respective objectives and answer the previously introduced research questions.

Design-based Research

Design-based research is a recognized methodology created by and for educators, that defined as "a series of approaches, with the intent of producing new theories, artifacts, and practices that account for and potentially impact learning and teaching in naturalistic settings" [BS04, p. 2]. It is also referred to as 'design research', 'development research' or 'formative research' and may provide valuable theoretical and practical contributions in educational technology research [Her+07; vdAkk+06]. The process of design-based research starts with an analysis of practical problems and continues with developing solutions using existing design principles and technological innovations. Next, iterative development and refinement cycles are conducted when applying the solution in practice. The complete process benefits from the collaboration of researchers and practitioners. Lastly, a reflection on the lessons learned will lead to new design principles and enhance the implementation of the solution. In each step, feedback for refinement of previous actions is collected.

Practical problems

As design-based research starts with an analysis of practical problems in a specific domain, this dissertation's starting point is somewhat similar. By reviewing existing learning games for security education and anti-phishing education, research gaps for developing different game prototypes are identified, and the practical problems of missing personalization and issues with unknown game content are identified through systematic literature reviews.

Iterative development

In the context of this work, the iterative nature of design-based research is implemented less strictly, as it might be intended and portrayed in other design-based research projects with multiple, more significant iterations. In the development of the game prototypes, a minor iteration using feedback from a preliminary evaluation is used. For the development of the personalization pipeline, the conceptual approach was refined during its implementation to fit the game requirements and provide modularity and extensibility for use in future work projects.

Collaboration between researchers and practitioners

Regarding the intended collaboration between researchers and practitioners in the design-based research approach, this work was supported by fellow researchers and colleagues of the Learning Technologies Research Group and the IT-Security Research Group at RWTH Aachen University, which can also be considered practitioners as they actively teach. Collaboration was also supported by publishing parts of this work and presenting it at different conferences of either the educational technology community or the security community, which usually offer a meeting point for researchers and practitioners in their domains. Furthermore, the complete research process was supported by the research training group 'Human Centered Systems Security' (sponsored by the state of North Rhine-Westphalia, Germany).

Results and lessons learned

As a result of design-based research, both lessons learned as well as design principles to enhance the implementation of the solution is expected. In this work, the evaluation of developed game prototypes, with or without personalization, yields valuable lessons learned and raises open questions and new future work directions to enhance the use of personalization in game-based anti-phishing education.

Experimental Hypothesis-testing Research

Experimental research is a research methodology widely recognized by various disciplines, such as medicine, psychology, and the social sciences. The key idea of experimental research is based on a design "utilizing randomized assignment of participants to conditions and systematic manipulation of variables with the objective of drawing causal inference" [Amec]. In the scope of this work, the experimental research design is used as part of *hypothesis-testing research*, a type of research whose purpose is to test a formulated research hypothesis and find causal relationships between different variables [Kot04]. As such, experimental hypothesis-testing research is based on the design of experimental setups in which participants are randomly assigned to different experimental conditions (as well as a control group) and evaluated to observe differences in response to initially stated hypotheses. Various instruments (e.g., questionnaires, tests) are used to evaluate different experimental conditions, and retrieved data is often analyzed using statistical methods. Expected results of experimental hypothesis-testing research include supporting evidence to accept or reject the different hypotheses and the creation of new hypotheses that should be tested in follow-up studies.

Types of variables

Differences between the experimental conditions or groups are quantified using independent variables. *Independent variables* are variables that are not dependent on any other variables in the study and can be manipulated by the researcher, e.g., different medication plans in the medicinal research. *Dependent variables*, on the other hand, are variables that are affected by the manipulation of independent variables, e.g., the participants' blood pressure or heart rate in medicinal research. By evaluating how the manipulation of independent variables influences dependent variables, researchers want to draw causal inferences to accept or reject the formulated research hypotheses. Essential for the evaluation of study results is the control of any other variables which are somewhat independent of the purpose of the study (also called *extraneous variable* [Kot04]). This way, the experimental error can be minimized, and effects upon the dependent variables can be attributed entirely to the independent variables.

Comparative studies with different games

In the scope of this dissertation, the objective in using experimental hypothesis-testing research is to evaluate different game prototypes with and without personalization and to understand how personalization may influence the participants' performance

in detecting phishing URLs. Comparative user studies are conducted, and different hypotheses are tested. While the different game prototypes serve as independent variables, participants' performance and confidence (measured using a test) serve as the dependent variables. The findings of statistical testings are used to answer study-specific research questions and the overall research question in this dissertation.

4.2 Methods and Tools

The use of methods and tools follows a mixed-method design, in which different methods are combined to fulfill the objectives of this dissertation This section briefly introduces different methods and tools in preparation for the following chapters.

Literature Review

In Chapter 5, literature reviews are used to achieve an overview of existing learning games in the domain of security education as well as the more specific subdomain of anti-phishing education. In the context of this dissertation, a *literature review* is defined as a part of the research process in which different sources of information in a specified research domain are compiled, described, and analyzed. Through a literature review, researchers draw connections between the sources and the research topic or question. Its purpose is to discuss the state of the art or relevant aspects of a specific research domain to gain input for new research.

Systematic literature review

Depending on the overall objective of the review, different types of literature reviews should be considered [GB09; Par+15]. The different types vary in their search, appraisal, synthesis, and analysis methods. As such, a *systematic literature review* "seeks to systematically search for, appraise and synthesize research evidence" [GB09, p. 95] and are often based on a specific guideline. However, these guidelines can vary between disciplines and are often not consistently used by different authors as they may or may not be suitable for the particular research topic. A very similar type to the systematic review is a *systematized review* which attempts to include elements of the systematic review but is less strict and formal in its search, appraisal, synthesis, and analysis [GB09].

Learning Game Design and Development

In Chapter 6, the findings of the literature reviews and identified research gaps of related work are used to guide the development of two new game prototypes. While there are various learning game design frameworks focusing on different aspects, the design of learning game prototypes in this dissertation are guided by Bloom's Revised Taxonomy [Kra02].

Bloom's Revised Taxonomy

Bloom's Revised Taxonomy (BRT) is a scheme for classifying educational goals, objectives and standards. The original taxonomy was published by Bloom et al. [Blo56] in 1956, and its revision was developed 45 years later by Anderson et al. [AKA01]. The taxonomy distinguishes between (1) the knowledge dimension and (2) the cognitive process dimension. For the knowledge dimension, four categories are considered: *factual*, *conceptual*, *procedural*, and *meta-cognitive* knowledge. The cognitive process dimension offers six different categories (similar to the original taxonomy): *remember*, *understand*, *apply*, *analyze*, *evaluate*, and *create*. The combination of both dimensions

forms the Taxonomy Table, which can be used to design and structure an educational intervention as well for the analysis of existing interventions to understand the learning goals and processes as well as evaluate curriculum alignment and missed educational opportunities [Kra02]. While BRT is used as a design aid for the development of new game prototypes (see Chapter 6.1), it is also used as an analytical tool for the literature review work presented in Chapter 5.3. Since the focus of this dissertation is not solely on game design and the development of new game prototypes was needed to evaluate the personalization of anti-phishing learning games, a particular future work direction includes to step back to the design phase using a proper design framework to improve the game prototypes further.

Multi-Touch Learning Game framework

For the development of the learning game prototypes, the *Multi-Touch Learning Game framework* (MTLG) is used. As introduced in Chapter 2.2.4, the MTLG framework is an open-source framework, which was initially developed for the design and implementation of collaborative multi-touch learning games for large scale multi-touch tabletop displays [ERS18]. However, the framework also supports game development for any other browser-capable device as it utilizes the HTML5 Canvas element and modern JS. In the scope of this work, the MTLG framework is used to implement the game prototypes described in Chapter 6. Furthermore, the current implementation of the personalization pipeline is also developed using JS and, as such, can be easily integrated into any game developed using the MTLG framework (see Chapter 7).

Study design

In Chapter 8, a user study following a pre-test/post-test between-group laboratory design was chosen to evaluate the developed game prototypes. For the evaluation of long-term effects and possible behavioral changes after playing an anti-phishing learning game, a longitudinal study design was chosen as an addition to the pre-test/post-test setup.

Pre-test/post-test design

A *pre-test/post-test* design is a study design in which the same assessment (e.g., a survey) is completed before and after an intervention or treatment [FH03; Amef]. In the scope of this work, the treatment or intervention is the act of playing either one of the learning game prototypes (described in Chapter 6). The comparisons of pre-test and post-test results are used to determine if any changes could be attributed to the intervention and the magnitude of these changes.

Between-group design

A *between-group* (or 'independent measures') design describes a study in which participants are divided into multiple groups or experimental conditions and thus, they receive different interventions or treatments [FH03; Amea]. Randomized participant allocation to the different experimental conditions is essential for between-group designs as it eliminates potential systematic effects and errors due to missing natural variation in the experimental conditions. Further details on between-group designs are described in [FH03].

Longitudinal design

A *longitudinal* design describes a study setup in which participants are tested in a series of assessments over a more extended period. It is also called interrupted time-series and includes multiple assessments scheduled around an intervention. One or more assessments can be taken before and after the intervention. In the scope of this work, the longitudinal design utilizes one additional measurement three months after the initial pre-test/post-test setup (see Chapter 8.3). Due to the interruptions by different periods between measurements, the testing is less controllable as variables could be influenced by various factors between measurements, e.g., more news reporting about phishing

attacks between two measurements could influence the measured perception of phishing as a risk.

Laboratory research is a study conducted either in a laboratory or in an environment where the investigator has some level of direct control over the environment and can manipulate the independent variables. (i.e., the intervention in a pre-test/post-test between-group design) [Amee]. Due to the COVID-19 pandemic, the comparative user studies and the longitudinal study were conducted in a remote lab study setup. The participants joined a video conferencing session using their personal devices. For playing the games and completing the surveys, participants were required to use a web browser. *Laboratory research*

Apparatus and Materials

For the evaluation of game prototypes in a user study, surveys were used to collect respective data about the participants. In the scope of this work, a *survey* is a set of questionnaires and tests used together in pre-, post-, or longitudinal tests. A *questionnaire* is a set of questions or prompts used to collect information from a participant, e.g., about a topic of interest or ability or other attributes [Ameg]. As such, a questionnaire regarding the participants' demographics is included in the surveys in this work. The complete surveys for pre-, post, and longitudinal tests were provided online using the software LimeSurvey[6]. In the following, different tests and questionnaires are briefly described to present an overview and prepare for the remaining chapters of this work. *Surveys and questionnaires*

To evaluate the effectiveness of the game prototypes in the scope of this work, performance tests are created to assess the participants' performance before and after playing the games. The *URL classification test* is a test that requires participants to classify URLs as either legitimate or phishing URLs. In addition to classifying a URL, participants are required to rate how confident they are in their decision on a 6-point Likert scale (from 1 = 'very uncertain' to 6 = 'very certain'). The test results include mean performance scores and mean confidence levels which can be further evaluated using statistical testing. Further details are provided in Chapter 8.1.2. *URL classification test*

Besides the participants' performance in classifying URLs, a questionnaire assessing the participants' perception of phishing as a risk is used. The questionnaire presents an adapted subset of the questionnaire used in [Ara12]. The questionnaire results in pre-and post-test allow for an evaluation of potential changes after playing the games. The adapted version contains 11 items using a 5-point Likert scale (from 1 = 'strongly disagree' to 5 = 'strongly agree') For details on how the questionnaire was adapted for the context of this work, see Chapter 8.1.2. *Perception of Phishing questionnaire*

A *Recognition of Services* questionnaire was used to assess the participants' familiarity with different services (e.g., 'eBay', 'Mobile.de'). The questionnaire consists of a set of services that the participants had to rate. The rating evaluated whether the participant uses a given service, only knows but does not use the service, or whether the service is unknown to the participant. For further details regarding the questionnaire, see Chapter 8.1.2. *Recognition of Services questionnaire*

For the evaluation of long-term effects and potential changes in behavior, a custom questionnaire called *Behavioral Change questionnaire*, was designed to assess the following four different constructs: (1) self-reported application of knowledge, (2) interest in learning more about security using games, (3) behavioral change and (4) the perception *Behavioral Change questionnaire*

[6] https://www.limesurvey.org/, accessed on 17.12.2021

of phishing as a threat (see Chapter 8.3.2). The questionnaire contains 9 items using a 6-point Likert scale (from 1 = 'strongly disagree' to 6 = 'strongly agree').

Statistical Analysis

Descriptive and inferential statistics

The results of previously mentioned questionnaires and tests were evaluated using statistical analysis (or testing), which is the application of different methods to describe, explore, understand, explain and test hypotheses about data [Ameh]. As such, two types of statistics were used in answering the research questions of different studies.

- *Descriptive statistics*, which is defined as the procedures for describing "the main aspects of sample data, without necessarily inferring to a larger population" [Ameb]. It is used to describe the participant samples in aspects like age, gender, level of education, and the performance scores and confidence levels in the URL classification tests.
- *Inferential statistics* (or inductive statistics), which is defined as a type of methods "that allow inferences about the characteristics of a population to be drawn from a sample of data from that population while controlling (at least partially) the extent to which errors of inference may be made" [Amed]. It is used to test different hypotheses using t-tests (e.g., by Welch or Student) or an ANOVA (Analysis of Variance).

Gameplay Analysis

Game Learning Analytics

Event log data is collected for each game session to understand what happens within a game. This data includes event statements of all actions performed by players and events related to players' paths through the game's structure. In the scope of this work, *gameplay analysis* (also referred to as Game Learning Analytics [Fre+16]) defines the process of analyzing event log data of different gameplay sessions by applying methods of data analysis. The objective is to provide insights into the gameplay sessions and provide detailed interactions about how players interact with the game [Fre+16]. As such, data pre-processing is used to prepare event log data for further processing steps as part of the different analyses. The methods for data analysis are always dependent on the game and what type of information is collected as log data, e.g., (descriptive) statistical methods using performance or timing data. For the gameplay analysis performed as part of this work, the focus lies on evaluating the players' performance and identifying mistakes or potential misconceptions.

5 Systematic Review of Existing Learning Games

This chapter covers two systematic reviews of literature and existing applications in the field of security education and anti-phishing education. By systematically reviewing the existing research and developments in the fields of game-based learning for security education and anti-phishing education, the current state of the art can be captured, and potential research gaps and open questions can be identified. In Chapter 5.1, related work on reviews and classifications in the domain of game-based security education is summarized. Next, a review of existing learning games for security education is presented in Chapter 5.2, while in Chapter 5.3, a review of anti-phishing learning games is described. Lastly, identified research gaps and open questions of both reviews are summarized and discussed regarding the overall scope of this dissertation in Chapter 5.4.

While large portions of this chapter were previously published in different publications (see [E], [O], [P] and [Q]), the content is presented in a more comprehensive, restructured way.

5.1 Existing Reviews and Classifications

In the past, various researchers reviewed different approaches to game-based security education [Alo+16; CEW15; HAB16; PDC10; TMJ17]. While some researchers focused on game-based learning approaches and compared them to traditional training and instruction, others compared games and different gamified concepts to summarize the state of the art at the time. These existing reviews emphasize the potential benefits, challenges, and problems of applying game-based learning in various forms. Besides reviewing existing work, researchers applied different classification schemes or methods to provide a systematic overview of games in security education.

In 2016, Alotaibi et al. published a review of existing learning games for security education and identified the use of gaming technology for IT security education as relatively new and in need of more extensive research [Alo+16]. Their review found games primarily targeting students and children and covering general security topics. No issue-specific games such as anti-phishing learning games have been reviewed. Further, Pastor et al. emphasized the potential of practical training and the use of simulations in game-based learning for IT security education [PDC10]. They reviewed 13 different simulation-based applications (with or without game characteristics) and applied a taxonomy for classification and further analyses. Next, Compte et al. presented different serious games and stated observations and suggestions regarding the games' design [CEW15]. They collected and reviewed academic publications and commercial products but did not elaborate on their methodology. In their work, Compte et al. concluded that reviewed serious games try to deliver an immersive experience through simulations and interactivity and also identified that most games are tied to formal learning contexts [CEW15].

In 2017, Tioh et al. argued for the potential of game-based learning to combine characteristics of both traditional and hands-on training methods [TMJ17]. Their review of published studies using available games showed that the effectiveness of game-based learning for IT security education was not yet empirically proven. Similarly, Hendrix et al. noted positive effects but criticized small sample sizes and non-discussed effect sizes in several studies [HAB16]. They also identified that while various games could be found in literature, the availability is limited as many projects were discontinued or never made publicly available.

Although various publications with reviews of game-based learning for security education exist in related work, no peer-reviewed publications in the research area of anti-phishing education were available to be included in this work. As existing reviews also look at games on other topics in the IT security domain, they are rather broad and less insightful on the specific topic of phishing. Therefore, there is a need for an overview and in-depth analysis of existing publications and available games in anti-phishing education. This analysis supports to focus on the specific characteristics of phishing, which stands out among IT security attacks in that it actively depends on user interaction and deception. Thus, two reviews on games in security education and anti-phishing education are presented in the following.

5.2 Review of Learning Games for Security Education

This section presents a systematic literature review of game-based learning for security education and takes a closer look at different, available learning games. In addition to the review, a game analysis using the G/P/S model [DAJ11] (described in Chapter 5.2.3) is used to provide a classification of retrieved learning games for the categories gameplay, purpose, and scope. This analysis builds on the results of a two-fold systematic literature and product review and leads to further insights on how learning games for security education are designed. The work was previously published in [P] and [O].

Research objectives

The research objective of this literature review is to provide an overview of existing learning games for security education targeted at end-users (as characterized in Chapter 1.1). As introduced in Chapter 2.1.2, end-users are users that have little or no prior formal education in CS and IT security. Furthermore, the objective is to find indicators that available games for end-users fail to teach sustainable skills and knowledge of CS and do not adequately educate their target group to behave securely and assess risks appropriately. Lastly, with applying the G/P/S model, another objective is to provide a systematic overview of the purpose and scope of existing learning games for security education.

Hypotheses

The following two hypotheses have been derived from the before mentioned research objectives:

- **H1**: There are not many game-based learning applications and learning games in the field of security education, which are targeted at end-users.
- **H2**: Available games for end-users do not teach sustainable skills and knowledge of CS to properly educate their target group to behave securely and assess risks appropriately.

The motivation for this review of learning games for security education follows the assumption that while a game has proven to be effective in scientific studies and for the scope of a research project, it seems to disappear after its evaluation. It, thus, fails

to reach its intended target group, e.g., end-users. The following sections cover the methodology, results, and discussion of the systematic literature and product review of learning games for security education, as well as the methodology and results of applying the G/P/S model [DAJ11].

5.2.1 Methodology

Since development and research in the domains of game-based learning and IT security are driven by academic and economic stakeholders, a systematic review of scientific literature may fail to cover all available learning games for security education. Thus, a two-fold review process entailing a systematic literature review on scientific publications and a product review was performed in September 2018.

Keyword selection

Both processes, the scientific literature review, and the product review are based on an initial search of which the result sets provide the input for the review. For the search, two keyword sets were used (see Table 1). These keyword sets contain the most suitable and important keywords in their respective categories but are not expected to be complete in coverage of all publications and products. While one keyword set contains security education-related terms, the other covers terms on game-based learning and learning games.

Table 1: Keyword sets used for systematic literature review

Set name	Keywords
ITsec	IT security, cybersecurity, risk awareness, security awareness, security education, cyber education, security
LearnTech	game based learning, gamification, serious game, learning game, edugame, teaching game, competence developing game

Search query

The search query was constructed as follows: $\{x \wedge y \mid x \in \text{ITsec}, y \in \text{LearnTech}\}$ (see Table 1 for the keyword sets). It was used in three different digital libraries, i.e., IEEE Xplore[7], Google Scholar[8], and ACM Digital Library[9]. For each search request, the first 100 results were extracted for review. The limit was set to 100 results since results with an even lower rank may fit the search query less.

Multi-step filtering and classification

Thereafter, a multiple-step filtering and classification process was performed to review all extracted publications from the result set systematically. First, duplicates in the result set were discarded. Similarly, all publications without online availability and accessibility (via university access or open access) were discarded since they could not be analyzed beyond their metadata (i.e., no access to full-text publication).

Next, all results were filtered based on the leading question, whether its topic is game-based security education or learning games for security education. All publications that were off-topic were discarded to reduce the result set. Based on this reduced set, a categorization was performed and results were sorted into the following categories: (1) game, (2) gamification, (3) competition, (4) review, and (5) others. The latter included all publications on frameworks, tools, implementations, and other security education content that does not fit into any category.

[7] https://ieeexplore.ieee.org/, accessed on 02.09.2019
[8] https://scholar.google.de/, accessed on 02.09.2019
[9] https://dl.acm.org/, accessed on 02.09.2019

Since reviews can be considered related work that take a look at learning games for security education, all cited publications and explicitly named learning games were added to the result set for further processing. This measure prevented missing games that other researchers already reviewed due to differences in the methodology.

To respond to the hypotheses derived from the research objectives, the next step included determining actual games addressed in the scientific publications of the result set. In addition, the covered security topics of the games and the intended target group and educational context were identified as best as possible. While some publications stated which target group and educational context a game was designed, these details were missing in other games and had to be estimated by checking the complete publication carefully.

Lastly, the online availability was determined. Since tabletop games, i.e., board games and card games, are most often non-digital games, online availability also refers to information available online.

Product search

Since scientific publications may not be available for games developed by companies and in industry, the result set was extended by performing a product search using the Google search engine[10]. This search was performed similarly to the search using the digital libraries. In addition, the result set of this search was checked for any games for security education. New and available games were added to the result set if not discovered via the scientific literature review. Also, the games' names, covered security topics, intended target group, and educational context was determined.

The final set of games was then used to respond to the hypotheses and provide results for the research objectives.

5.2.2 Results

After running all search queries in the before-mentioned digital libraries, the initial results contained 2636 publications. By discarding duplicates and inaccessible publications (i.e., publications for which no access to the full text was available), the set was reduced to 1277 results. Next, the results were filtered based on the content, i.e., game-based security education by any means. The remaining set contained 183 publications.

Categorization of types

Next, each publication in the result set was categorized by type. As such, 14 papers were identified to be publications of different literature reviews comparing existing games for security education. By adding the games mentioned in these review papers to the result set, it was extended to 216 results. In this final set, 181 results are of type 'game', 'gamification' or 'competition'. A cumulative overview is shown in Table 2. For further processing, only publications categorized as 'game', 'gamification', or 'competition' were considered. Papers marked as 'Others' were discarded as they are not about any game for security education.

Exclusion of results on competitions

Publications categorized as competitions may include Capture the Flag competitions (CTFs), hackathons, or other security challenges. Although they may include game elements, they differ from game-based learning applications and learning games, as they are no game by design. Also, the search query was not designed to discover publications on competitions like CTFs, hackathons, or similar events; thus, the result set may not cover a representative portion of available competitions. The keyword set *LearnTech* was focused on game-based learning and learning games rather than on competitions and challenges (see Table 1). After reviewing all results categorized as competitions, about

[10] https://www.google.de, accessed on 02.09.2019

Table 2: Cumulative overview of the result set

Type	Number of results
Game	133
Gamification	24
Competition	24
	$\sum = 181$
Review	14
Others	21

two-thirds of the addresses competitions targeted CS students or professionals. Thus, it is to be expected that they are unsuitable to end-users. While the remaining third of the competitions might be suitable to end-users, they are most likely for participants who are highly interested in the topic and have possible prior knowledge. Conclusively, the results categorized as competitions were discarded from the result set since they are no games and not primarily for end-users without prior knowledge in CS and IT Security.

Exclusion of results on gamification

Next, all results categorized as gamification were discarded. Since gamification and games differ in their (re-)design of the learning task, publications on gamification in security education were not considered further in this review. The difference between gamification and game-based learning is described further in Chapter 2.2.1.

Analysis of target groups

To this end, the following analysis will based on the set of 133 results categorized as games. As described in Chapter 5.2.1, the next step of the analysis included the determination of the games' names, covered topics, intended target groups, and intended educational contexts. The analysis revealed 99 different learning games. Possible target groups included:

- students (either in the field of CS or not),
- employees (either professionals in IT or general employees),
- end-users, and
- others (e.g., parents).

Analysis of online availability

While 99 different games seem to be a lot, the online availability of these games is a crucial aspect for further interpretation. Digital games which are not available online (either as browser games or downloadable applications) are inaccessible to the intended target group. Online availability is less critical for non-digital games, board games, and card games. It is rather important whether those games are still on the market and available for purchase or through other channels. In some cases, even non-digital games can be found online, e.g., as downloads that can be printed. For the set of games in the scope of this review, only 48 out of 99 games were available online. With this included are also all non-digital games still available through other channels. Results of both, target groups and online availability are summarized in Table 3.

Analysis of educational contexts

Besides the target groups of the games, the intended educational contexts have been determined. The following contexts have been considered: (1) Primary school, (2) middle school, (3) high school, (4) college/university, as well as (5) corporate and (6) non-formal contexts. Since some games turned out to be intended or suitable for multiple different educational contexts, multi-label classification was applied (see Table 4).

Table 3: Distribution of Target Groups and Availability

Target group	# games	# available games
CS students	19	5
Non-CS students	32	18
Professionals	9	5
Employees	12	7
End-users	26	12
Others	1	1

Table 4: Results of educational context analysis using multi-label classification on all games

Educational context	# games
Primary school	4
Middle school	9
High school	10
College/university	26
Corporate	20
Non-formal	38

Among the identified games, the most common educational contexts are 'college/university', 'corporate', and 'non-formal' contexts. Since learning games for security education would most likely be used in CS classes, it is not surprising that fewer games are intended for primary or middle school. In most countries, CS education is not covered in primary and middle school curricula; therefore, fewer games are explicitly designed for those school contexts. It can be assumed that security education will not be part of the curriculum without CS education.

The situation might differ for high school curricula or study programs at the college/university level. CS education coverage can be assumed higher for these contexts, and thus, more games are designed for these target groups. In addition, games that are part of a research project are usually tested. It is not uncommon to utilize students to test and evaluate the games. While these games might be integrated into the curriculum of the respective class they are used in, they could also be suitable for end-users in non-formal learning contexts depending on the required prior knowledge.

Regarding the corporate learning context, games are often developed professionally and used for the training purposes of employees. The embedding into the corporate context can make games less suitable for private end-users, e.g., they simulate a corporate environment to create an authentic learning experience, and company-specific aspects are integrated into the game. Topics may include corporate espionage or data privacy regulations of a company. It is questionable whether these games are suitable for end-users since it might be harder to relate the games' content to the end-users' private, day-to-day activities.

Analysis of topics

The analysis of topics covered in the games did not yield clear results but rather a set of reoccurring issues in different games, e.g., phishing and password security. Ranging from network security to hacking and online safety, available games often cover more than one topic and differ in the level of depth and detail.

In conclusion, the results of the two-fold approach to reviewing learning games for security education consist of a set of 48 available games. While some games might be

suitable for multiple educational contexts, they are usually designed for one particular target group. The content covered in these games varies, whether the topic or the level of detail.

5.2.3 Game Analysis using the G/P/S model

As an extension of the systematic literature and product review on learning games for security education, a classification scheme for serious game is applied to the set of available games. Parts of this work were previously published in [O] and [Q].

Methodology

The G/P/S model is a classification scheme for serious games first presented by Djaouti et al. [DAJ11]. It was designed to solve problems of previously developed models and taxonomies for serious games and focuses on the analysis of the two dimensions behind the term *serious game*. The model is a multidimensional classification scheme that addresses the 'serious' dimension as well as the 'game' dimension by classifying games in three aspects: (1) gameplay, (2) purpose, and (3) scope [DAJ11].

Gameplay

The 'gameplay' aspect captures how the game is played. Therefore, details about the game structure need to be determined. While there is no uniform definition of gameplay in academia and the gaming industry, the concept explains how the game is played and the general objectives of the game. In [DAJ11], the gameplay is distinguished according to a game's objectives: Games without rules checking for the objectives are classified as 'play-based', while games with rules that define the game's objectives (e.g., winning or losing) are classified as 'game-based'.

Purpose

For the 'purpose' aspect, Djaouti et al. distinguish between three main purposes of a game: (a) message-broadcasting, (b) training, and (c) data exchange [DAJ11]. While learning games often convey knowledge or foster content understanding and include possibilities to practice, their purpose is message-broadcasting or training. Collaborative games in which the interaction between players determines success in the game can have the purpose of data exchange, e.g., Lure of the Labyrinth.

Scope

Lastly, 'scope' refers to the area of application and the intended audience (e.g., end-users, students, professionals) and market of a game (e.g., education, military). As such, it describes the target group of a game.

For the context of this review, the G/P/S model is used for the classification of identified games in the result set [DAJ11]. Each game in the result set is classified in all three aspects to enrich the results of the review and provide input to the discussion in Chapter 5.2.4.

Results

The game analysis using the G/P/S model [DAJ11] was performed on 39 available games from the result set of the prior review work. It should be noted that nine games of the original result set were no longer available at the time of the game analysis (March 2019). Unavailable games were not considered for the analysis since playing the game was required to determine classification results in all three aspects of the G/P/S model. The results of the game analysis are summarized in Table 5.

Table 5: Results of game analysis using the G/P/S model. For 'purpose' and 'market', the assignment of multiple classes was possible.

Gameplay	# Results
Game-based	17
Play-based	22
Purpose	**# Results**
Message broadcasting	27
Training	15
Data exchange	4
Scope	**# Results**
Market	
Education	30
Corporate	8
Entertainment	1
Target Group	
General Public	20
Professionals	8
Students	18

Analysis of Gameplay

The aspect of 'gameplay' was determined by identifying the objectives and rules of a game. Games with objectives and rules for winning are classified as 'game-based', while games without explicit rules for fulfilling the game's objective (e.g., rules for winning) are classified as 'play-based'. The latter often contains game elements like rankings, scores, and leaderboards since the game's main objective may be to achieve high performance. Among the 39 available games, 17 games are classified as 'game-based', while 22 are classified as 'play-based'.

Analysis of Purpose

Since a game can serve multiple purposes, the assignment of multiple purposes was allowed. The most occurring purpose was message broadcasting (27 of 39 games). Only 15 games offer training opportunities, e.g., in the classification of emails where players have to decide whether an email is malicious or benign. Only four of the available games allow for collaboration and data exchange. It should be noted that these four games are all non-digital games, i.e., card games or board games. Here, collaboration through discussion and social interactions activates the exchange of ideas, making joint decisions, or assessing risks.

Analysis of Scope

Regarding the scope of the games, a distinction was made between markets and target groups. The majority of the games were developed for the educational market. 8 of 39 games are intended for the corporate environment, while only one game was created for the entertainment sector. The target group distribution shows that more than half of the analyzed games are suitable for the general public. The second most targeted audience is students, which aligns with the market distribution. Like the results of the market classification, 8 of 39 games were intended for professionals.

In conclusion, the game analysis using the G/P/S model [DAJ11] provided a systematic overview of available games for security education. While the analysis of gameplay results indicates the use of both game-based and play-based approaches, most games

have the purpose of message broadcasting or training, which aligns with the most addressed market, i.e., education. Therefore, it is not surprising that many games are for students and the general public. Both are common target groups of security education as explained in Chapter 2.1.2.

5.2.4 Discussion

In this section, the results of the two-fold systematic literature and product review and additional game analysis using the G/P/S model [DAJ11] are discussed, and the initially formulated hypotheses and research objectives (described in Chapter 5.2) are revisited.

In the first step, the presented review resulted in a set of 99 games, out of which 48 were publicly available at the time of the review. In the later game analysis with the G/P/S model, only 39 games were still available. Nine games went offline or were not accessible anymore. This result supports the initial assumption that although games have been developed, researched, and potentially proven effective or suitable for the intended target group, they seem to disappear over time. Lastly, they do not reach the target groups in the long run.

Hypothesis H1

Regarding the first hypothesis (**H1**), the goal was to review existing game-based learning applications and learning games in the field of security education and assess whether or not they are intended for end-users. The review results show that more than half of the games are designed for the respective target group. Among the 58 of 99 games, 26 games are suitable for end-users and 32 for Non-CS students. As described earlier, games for employees might not be suitable for end-users and, thus, they are not considered in response to the hypothesis. Of these 58 games, only 30 games were available at the time of the review. While at first, these results indicate that many games are intended for end-users without prior knowledge in IT security and CS, the fact that only about 50% of them were still available weakens the result.

The game analysis using the G/P/S model supports these observations. Among the 39 analyzed games, 20 games are suitable for the general public (see Table 5). In addition, 18 games are suitable for students and could be suitable for end-users depending on the games' required prior knowledge and focus. It should be noted that the G/P/S model does not distinguish between CS and non-CS students. Thus, to include these games would require a more in-depth analysis of the games' content.

According to the results of the game analysis, the most targeted market is the educational market (76.92%), which includes formal educational contexts, e.g., schools and universities, and non-formal contexts. This is no surprise since the G/P/S model was initially intended for serious games and learning games are a type of serious games with educational content. Among the rest, 20.51% are suitable for the corporate market, and only one game is intended for entertainment (2.56%). Unfortunately, the market analysis does not yield any information about necessary prior knowledge among the targeted market. It could be argued that for the entertainment market, games should not require prior knowledge to appeal to large portions of the market. However, only a closer look into the games and the covered content may support this argumentation.

Conclusively, the results of both the literature review and the classification using the G/P/S model support the hypothesis that there are not many game-based learning applications and learning games in the field of security education, which are targeted at end-users (**H1**). While there are numerous games suitable for end-users in the literature,

the actual availability limits the games' effectiveness in reaching the target group of end-users. Among 39 analyzed games and a total of 99 games mentioned in the literature, only 20 games are suitable for end-users. As such, hypothesis **H1** is accepted.

Hypothesis H2

For the second hypothesis (**H2**), the goal was to evaluate whether available games for non-professional end-users teach sustainable skills and knowledge of CS to properly educate their target group to behave securely and assess risks appropriately. Therefore, it is necessary to determine what available games teach and how knowledge and skills are conveyed. This was done by reviewing different games in the set of available games, including the respective literature about these games. The overall content analysis of the games turned out to be rather difficult, since topics in security education are approached on different levels of detail and thus, they are difficult to compare. In the following, four different games are analyzed as examples to provide indicators regarding the second hypothesis.

Figure 1: Screenshot of "The Internet Safety Game" which shows the board game structure where the player's character can move along pathways and collect items.

The Internet Safety Game

First, a game called 'The Internet Safety Game' was reviewed. It is a browser game which was available on the platform 'NetSmartKidz'[11] by the National Center for Missing & Exploited Children[12] at the time of the review. Its target group is younger children in non-formal learning contexts. The gameplay is similar to a board game, where the board consists of pathways, and a character can be moved step-wise along these pathways when rolling dice. The player has to collect various items on the board to advance. After collecting up to six items, the player must complete a multiple-choice quiz as a final assessment. Depending on the chosen difficulty level at the beginning of the game, the final assessment is skipped, and players win by collecting all six items. Each item is a piece of textual information regarding the Internet and online safety. The information shared with the collectible items is structured as recommendations on suitable, safe online behavior or Internet vocabulary. While the recommendation of not sharing personal information seems to be appropriate advice for younger children, the advice is entirely out of context. There is no explanation of potential risks and reasons why someone should not share personal information online. In addition to the missing context, the game's content is very limited since it only requires the player to collect six items, and these are only short textual pieces of information.

[11] https://www.netsmartzkids.org/, accessed on 02.09.2019
[12] https://www.missingkids.org/HOME, accessed on 02.09.2019

PASDJO

Next, the game prototype PASDJO by Seitz and Hussmann [SH17] was reviewed. In this game, the players learn about password security by rating different passwords and receiving feedback on the quality of passwords accordingly. The gameplay is kept short and straightforward. Although utilizing feedback on the players' ratings to present learning content, the game does not take it further. Topics related to password security are not addressed, e.g., risks of weak passwords, authentication mechanisms complementing passwords, or attack schemes to break passwords. Like 'The Internet Safety Game', the learning content in PASDJO is presented entirely out of context. In addition, Seitz and Hussmann do not provide any description of learning goals [SH17]. Regarding the second hypothesis, this game could be considered a game that fails to teach sustainable skills and knowledge since it is very limited in its learning content.

Safe Online Surfing

The 'Safe Online Surfing' platform[13] by the Federal Bureau of Investigation (FBI) provides a series of mini-games for children of different age groups (sorted by grade levels in school). The mini-games incorporate different game elements and mechanics to engage players. However, similar to the first two games, the available mini-games do not provide proper context and cover rather selective factual knowledge. Therefore, they are not considered suitable to teach sustainable skills and knowledge.

CyberCIEGE

The fourth and last game considered for the discussion of this review is CyberCIEGE, which was developed as a research prototype by the Naval Postgraduate School and Rivermind [ITA05]. The game focuses on computer and network security. It provides an interactive learning environment in which players take on the role of employees responsible for firewalls, Trojan horses, and other security-related infrastructures of an IT-dependent organization. The game's objective is for players to provide necessary security measures to protect the organization's assets while its virtual users are undisturbed and productive. CyberCIEGE challenges players to build and configure computer networks, and the players' choices affect the virtual users' productivity as well as the attackers' abilities to compromise the organization [ITA05]. The game provides simulation-based scenarios on different IT-security topics, e.g., malware. Compared to the previous three examples, CyberCIEGE provides learning content embedded in a specific context. The underlying narrative of an employee configuring IT systems and dealing with security-related issues is derived from a real-world setting and provides an immediate context for players. However, CyberCIEGE was developed for formal learning contexts, usually supported by an instructor or teacher. Therefore, a notional syllabus[14] is provided and the game's effectiveness in educational contexts has been evaluated in various user studies [AAS16; RLA14; Jon+10; Fun+08; ITA05]. While CyberCIEGE may provide suitable learning content on security-related topics and deliver it in a relatable context, its primary target group is students in formal learning contexts. Thus, the suitability for end-users is questionable.

The presented games (and game platforms) are only a subset of games identified in the literature review. They show some limitations regarding the selection of content and requirement for context. For example, while CyberCIEGE is rich in content and covers various topics in a context, the other games are rather selective and present learning content entirely out of context. However, while end-users could play CyberCIEGE, it was designed for formal learning contexts and came with a curriculum, which explains its wide range of content. The other games and platforms are freely available to the end-users and only present short mini-games that can be played without prior knowledge. Their content varies and seems somewhat arbitrary and without context, however since

[13] https://sos.fbi.gov/en/, accessed on 02.09.2019
[14] https://nps.edu/web/c3o/syllabus, accessed on 07.05.2021

'The Internet Safety Game' and the 'Safe Online Surfing' platform are designed for younger end-users, the content selection might differ from a game for adult end-users. To this end, a scoping review regarding one particular target group, e.g., children, would be worth exploring. By reviewing games for a particular target group and comparing game content with relevant learning content, e.g., extracted from curricula for this target group, the quality of the game could be assessed. This way, the hypothesis of sustainability of learning content could be answered under consideration of a specific target group and demographic.

Another approach to assessing learning content quality would be to focus the complete literature review on a subdomain of security education, e.g., anti-phishing education. By selecting a more narrow content domain (e.g., phishing), learning game content could be compared on a more detailed level. Here, a domain expert could be consulted to analyze the games' content and compare it to necessary learning content in the particular subdomain of security education. This could also be paired with the analysis of learning goals, e.g., using BRT, and evaluating the learning goals with respect to addressed cognitive processes and types of conveyed knowledge.

In response to the second research hypothesis (**H2**), no clear decision was made to either reject or accept the hypothesis. While the learning content of reviewed games seems rather selective and somewhat arbitrary in some cases, games like CyberCIEGE show that content-rich games might be designed for formal education and less suitable for end-users. Another reoccurring issue in some games is the problem of context, i.e., learning content is embedded into the game but without any effort to place it into a suitable context. Games like 'The Internet Safety Game' allow players to collect information pieces, but each piece is loose and without context information. As such, the question regarding the sustainability of learning game content is yet to be answered.

In the scope of this dissertation, a follow-up review is presented in the next section. This literature review focuses on anti-phishing learning games, i.e., learning games about a subdomain of security education. As suggested, the review focuses on a particular topic to allow for a more comprehensive analysis of the learning game content and the learning goals.

5.3 Review of Learning Games for Anti-Phishing Education

As a follow-up work to the systematic review of learning games for security education, this section presents another review focusing on anti-phishing learning games as one particular sub-genre of learning games in security education. This work follows a three-step approach to reviewing existing publications and available games. Based on a derived data set of publication on games, the in-depth analysis encompasses the identification of target groups, educational contexts, learning goals based on BRT [Kra02], and covered learning content. This work was a collaboration with Klemens Köhler and Vincent Drury, where everyone collaborated equally. It was previously published in [E]. Furthermore, it extends methodology used in the previous review as well as in [C].

Research objective statements

In order to find out how effective available anti-phishing games can be at preparing end-users to defend against phishing attacks, this review is guided by the following research objective statements:

- (RO 1) A literature review was conducted to create a comprehensive data set and answer which anti-phishing games are available and what is their intended educational context and target group.

- (RO 2) A further analysis of the literature, using BRT for categorization, was then done to establish whether the knowledge and skills conveyed through the game mechanics are suitable and sufficient to prepare end-users.
- (RO 3) By playing the accessible games, extracting their content, and comparing those to contemporary phenomena in phishing attacks, the correctness and completeness of the conveyed knowledge was assessed.

As introduced in Chapter 2.1.2, end-users are users that have little or no prior formal education in CS and IT security. It can also be assumed that they have no access to organizational resources, such as training programs or immediate IT support. Consequently, learning games with end-users as their target group have to be effective in informal learning contexts and provide a safe environment for end-users to experiment and learn.

In order for learning games to be effective, they have to provide suitable learning content with the appropriate methods. Therefore, instructional principles and methods can be mapped to game mechanics [Arn+15]. Existing research has shown that while a game's learning content is often described in available publications, insights into the methods and game mechanics are not [C].

Hypotheses

Next, the research objectives are translated into the following hypotheses:

- **H1**: The majority of games are designed for end-users without prior knowledge in IT security.
- **H2**: The majority of games are designed for informal learning contexts.
- **H3**: The game mechanics aim at specific, shared learning goals, and:
  - **H3a**: There is no difference in learning goals between the mechanics of digital and non-digital games.
  - **H3b**: There is no difference in learning goals between the mechanics of games that exclusively cover phishing and games that cover additional forms of IT attacks and defense.
- **H4**: The games require learners to utilize procedural knowledge.
- **H5**: The games convey detailed conceptual knowledge.
- **H6**: The games do not contain knowledge on advanced contemporary attacks.

5.3.1 Methodology for Data Set Creation

This section presents the methodology used for the literature and product search. Similar to the methodology described in the previous section, this work is based on the method of systematic literature review (see Chapter 4). While in the previous review, the focus was on learning games for IT security education in general, this review focuses on the collection and analysis of existing research in the field of learning games for anti-phishing education.

Keyword selection

At first, different digital libraries were queried (IEEE Xplore[15], ACM Digital Library[16], and Google Scholar[17]). All search queries were constructed by a combination of the keyword *phishing* and one keyword k from the following keyword set using the

[15] https://ieeexplore.ieee.org/, accessed on 13.07.2020
[16] https://dl.acm.org/, accessed on 13.07.2020
[17] https://scholar.google.de/, accessed on 13.07.2020

logical conjunction operator ($\wedge$ or AND). The collection of search results was limited to publications written in English or German. The initial result set contained 497 publications, and after discarding duplicates, the set was reduced to 282 results.

$$k \in \{ \text{educational game, serious game, learning game, game based learning, competence developing game} \}$$

Creation of POG data set

Next, the title and abstract of search results were analyzed for relevance to the research domain, i.e., learning games for anti-phishing education. This reduced the result set to 61 publications. Among these 61 publications, seven reviews on learning games were found. After checking each review for publications on learning games for anti-phishing education, the review publications were excluded. The reviews did not yield any new publications on anti-phishing games, which were not already part of the result set. Lastly, the results were analyzed to identify unique games described in the publications. The final *Publications on Games* data set (POG) contains 54 results, describing 40 unique games. For games described in more than one publication, only one publication was referenced in Appendix B.1. In the following sections, an in-depth analysis focusing on (1) target groups and educational contexts as well as (2) learning goals based on BRT and (3) the learning covered content is presented.

5.3.2 Analysis of Target Groups and Educational Contexts

Methodology

Depending on the learning content and specific learning outcomes, learning games are usually developed for a particular target group and are intended to be used in a specific educational context. While it is not expected that both pieces of information are directly revealed in the games themselves, a review of available publications on the learning games can be used to identify the target group and intended educational context.

Target groups

Similar to the previous review work on games for security education, it can be distinguish between the following target groups: (1) *CS students*, (2) *non-CS students*, (3) *employees*, (4) *IT employees*, and (5) *end-users*. If no target group is mentioned, the label (6) *unspecified* is assigned. Also, it can be explicitly differentiated between CS students and non-CS students since prior knowledge may vary significantly between these target groups. Similarly, a distinction is made between employees and IT employees. The underlying assumption is that tailoring a game towards a target group with prior knowledge of CS and IT security means building upon it. Thus, it might not be suitable for target groups without prior knowledge and should be distinguished from games that do not require knowledge.

Educational contexts

For the educational context, different levels of education are considered: (a) *primary school*, (b) *middle school*, (c) *high school* and (d) *college/university*. Further, (e) *corporate* and (f) *informal* learning contexts are considered. Lastly, unknown educational contexts are classified as (g) *unspecified*. The school contexts, college/university, and corporate contexts can be summarized as formal learning contexts.

Results

Among the 40 identified games of the POG data set, it was possible to identify 34 games (85%) with an explicit target group and educational context. The results for the

analysis of target groups and educational contexts are summarized in Table 6. It should be noted that the assignment of multiple classes for a game was allowed since games could be intentionally designed for multiple target groups and educational contexts (see Appendix B.1 for the complete analyses results).

Table 6: Analysis results for target groups and educational contexts

Target group	**# games**	**Educational context**	**# games**
CS students	4	Primary school	3
Non-CS students	9	Middle school	5
Employees	6	High school	1
IT employees	3	College/University	7
End-users	13	Corporate	9
unspecified	6	Informal	14
		unspecified	6

For the remaining six games (15%), in none of the publications, any target groups and educational contexts were specified. Thus, an educated guess was made based on the details presented in the publications, e.g., game description, screenshots, and participants of user studies. All six games are classified as suitable for end-users in informal learning contexts.

Hypotheses H1 and H2

In conclusion, the analysis results show that only a few games are designed for CS students and IT employees. End-users (13), non-CS students (9), and employees (6) are the most addressed target group among the analyzed games. In addition, no explicit notions of required prior knowledge for those games were found during the analysis, which supports **H1**. In addition, 14 games are suitable for informal learning contexts, supporting the second hypothesis (**H2**).

5.3.3 Analysis of Learning Goals and Taxonomy Levels

Methodology

This part of the analysis aims at the games' learning goals and the knowledge and skills conveyed using specific game mechanics (as introduced in Chapter 2.2.1). To classify the knowledge and skills acquired through playing the game, BRT is used [Kra02]. As introduced in Chapter 4, the taxonomy defines activities as learning goals along a cognitive process dimension and a knowledge dimension. It is depicted as a discrete Cartesian coordinate system, where each combination of a cognitive process on one axis and a knowledge type on the other constitutes a category. More detailed information on the concept and use of the BRT can be found in [Kra02].

The analysis was conducted on the academic publications in the POG data set, assuming that publications on games provide reasoning and reflection on game design choices. Thus, 40 unique games could be analyzed, and in each case, the sources contained information on game mechanics.

Modified semantic analysis

For data collection, a modified approach similar to the one used in [C] is used. Instead of analyzing how authors define the learning goals of their games, it was analyzed how they describe game mechanics. The original semantic analysis in [C] extracted learning goals by searching academic publications for indicator phrases and verifying

that they were used to describe learning goals. However, this approach has two drawbacks: Firstly, it relies on authors using the same or similar vocabulary to describe their goals, and secondly, it relies on the authors understanding of how to translate the learning goals in their game design. For example, many learning games in the analysis take a multiple-choice approach. While these games often include text elements to convey factual, conceptual, and procedural knowledge, the game mechanics themselves do not require using it. Facts like what constitutes phishing (fact) and how phishing URLs can be distinguished from legitimate URLs (concepts) are represented and must be remembered and understood (cognitive processes) to solve a quiz and advance in the game. A question might be: Where do you find the top-level domain in the URL? Here, only remembering and understanding conceptual knowledge is necessary to answer the question type, and it does not constitute the application of procedural knowledge. Furthermore, learners do not actively apply, analyze, evaluate, or create knowledge in the sense of the BRT.

Notation of results

By refining the method to analyze the descriptions of game mechanics, the actual activities required of the learners can be understood. If these in-game activities match activities defined in the BRT, specific categories of the knowledge dimension and the cognitive process dimension can be assigned. Multiple categories can be assigned to each game. As a result, each game had its own matrix of the BRT, where each category contained either a 1 (representing yes) or 0 value (representing no). For detailed examination according to **H3a** and **H3b**, two additional characteristics of each game were included: whether they are (a) *digital* or (b) *non-digital*, and whether (1) *yes*, they are exclusively about phishing, or (2) *no*, they cover additional IT attacks.

Results

The general overview of the results reveals that most games cover factual and conceptual knowledge to remember and understand. All other categories contain 15 games or less (37.5 %). The abundance of each cognitive learning goal tends to decline, going toward more complex cognitive processes and more abstract knowledge. Table 7 presents a summary of the results, while Table 33 in Appendix B.1 contains all detailed results of the analysis.

Table 7: Number of games covering BRT categories ($n = 40$)

	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual Knowledge	34	27	14	10	5	2
Conceptual Knowledge	35	23	10	4	4	1
Procedural Knowledge	15	11	13	5	6	3
Meta-cognitive Knowledge	6	3	5	2	2	1

Among the 40 unique games, 15 games focus solely on anti-phishing education without addressing further topics. These games concentrate even more on remembering and understanding facts and concepts (see Table 8). No game requires learners to evaluate or create knowledge, and only one game conveys meta-cognitive knowledge. The most used game mechanic is a binary decision mechanic in which learners have to classify either a URL or email as benign and phishing.

Table 8: Number of anti-phishing learning games in each BRT category ($n = 15$)

	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual Knowledge	15	9	4	2	0	0
Conceptual Knowledge	15	8	2	1	0	0
Procedural Knowledge	6	3	5	0	0	0
Meta-cognitive Knowledge	1	0	0	0	0	0

For further analyses, the games were divided into digital (31) and non-digital games (9). Looking only at the 31 digital games, there is again a complete lack of games requiring learners to create knowledge and a severe lack of meta-cognitive activities (see Table 9).

Table 9: Number of digital games covering BRT categories ($n = 31$)

	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual Knowledge	29	22	9	5	2	0
Conceptual Knowledge	30	19	5	2	2	0
Procedural Knowledge	12	6	11	1	1	0
Meta-cognitive Knowledge	2	1	0	0	0	0

Hypothesis H3

In conclusion, the results support the acceptance of **H3** since the vast majority of all games require remembering facts and concepts. A majority of the games also require understanding facts and concepts. Generally, as one advances along both axes of the BRT toward more complex learning goals, the number of games that convey them declines.

Hypotheses H3a, H3b, and H4

The results do not support the acceptance of **H3a**, **H3b**, and **H4**. **H3a** and **H3b** - the hypotheses that digital games and games exclusively about phishing attacks have similar learning goals as the complete sample - have to be rejected. Although the pattern mentioned above can be found no matter which sub-sample is considered, no digital game asks the learner to create, and no game exclusively about phishing asks the learner to create or evaluate. Additionally, digital and pure anti-phishing games do not cover most activities that require meta-cognitive knowledge. Finally, **H4** has to be rejected, as remembering procedural knowledge is only required in 15 games, and other cognitive processes in combination with procedural knowledge are necessary in even fewer games.

5.3.4 Analysis of Learning Content

Methodology

The results of the previous section show that many games require factual and conceptual remembering and understanding to advance within the games. In this section, the analyses presented are extended by examining the specific topics the games present and teach during a typical playing session. This analysis reveals which subjects are covered in current anti-phishing games, the level of detail they are presented, and whether specific topics are missing even if a broader subject is included in the game. To this end, the content of available digital learning games is reviewed and analyzed based on a number of subjects and topics. First, the 9 available games described by publications in the POG data set (cf. Appendix B.1) were analyzed as well as the collection by 4 reference games without academic publications that were found using the Google search engine[18] (9+4 = 13 games, see Table 10). These reference games represent games about phishing emails, URLs, and websites as offered by several companies.

Phishing-related subjects

First, the subjects concerned with phishing and covered in the games were identified. The most common subjects of the analyzed games were:

- *URLs and Websites* (main subject of 4 games),
- *Emails* (main subject of 4 games), and
- *Other/Various* (5 games).

Games in the *Other/Various* category do not focus on phishing but instead include learning content about phishing as well as several other topics of security. In the next step, more specific topics for each subject were defined, again based on the actual content of the games, while also including topics derived from research on common and advanced phishing attack techniques.

For URLs, topics were defined based on (1) the structure of URLs, (2) types of deceptive content, and (3) advanced topics. Regarding (1), it has been shown that attackers can make use of different parts of the URL to insert deceptive keywords [MG08]. The types of deceptive content (2) include common transformations in phishing URLs as well as advanced techniques like abusing Internationalized Domain Names (IDN) [ES18]. Lastly, for (3), a selection of URL types was reviewed, i.e., URL types that users are likely to encounter in their daily browsing or specific phishing attacks, like redirection, link-shortening, or services that host user-generated content (e.g., cloud services).

For emails, the following features were considered: (a) specific traits, (b) sender spoofing, (c) email structure, and (d) email attachments. Specific traits (a) are a common topic in games, even though they can sometimes be easily avoided by attackers. There are also several types of sender spoofing (b), with differences in the display-name and sender address [Res08]. Finally, knowledge about the email structure (c), e.g., email headers, can often be used to detect anomalies in emails [HW18]. Games that focus on emails often also include some examples or lessons on domain names, as they commonly appear in email headers and are also presented in the user interface of popular email clients.

Lastly, auxiliary topics were considered, including different attack channels (e.g., SMS, social media), advanced protection strategies (e.g., multi-factor authentication (MFA)), and common traits of the body of phishing websites.

[18] https://www.google.de, accessed on 07.07.2021

Analysis through playing

For the analysis, games are processed as follows: First, all available digital games were accessed (and downloaded if needed). It should be noted that only games that are generally available online, requiring no payment, membership, or request for access, were considered. Next, the games were analyzed through playing (i.e., from a player's perspective) while keeping note of the subjects and specific topics covered in the games. Then, each specific topic was rated using one of four different levels:

- 0 - *topic does not appear*,
- 1 - *topic does appear in game elements or examples*,
- 2 - *topic mentioned but is not fully explained*, and
- 3 - *topic is fully explained*

. For these distinctions, it was assumed that detailed explanations (level 3) are more likely to convey a detailed understanding of the actual detection or protection strategy, while shallow descriptions (level 2) can be confusing and might lead to misunderstandings (e.g., [She+07]). However, it can be argued that these beforementioned instances of levels (2) or (3) are more likely to be actively considered by learners than seemingly accidental information that may be hidden in examples or game elements (level 1).

The approach is limited to encountered gameplay in the analysis, but the playing sessions were especially focused on phishing when presented with a choice in the games. It should be noted that no claim is made to have encountered all examples or exhausted all possible selections, successes, and failures, though it was tried to cover all content related to phishing, or if the game's focus is phishing, complete one gaming session.

Results

The availability for all digital games in the POG data set was determined by searching the corresponding publications, querying a search engine, as well as searching popular code repositories like GitHub[19] for references to the game. This way, 13 games were obtained, which includes the four reference games that were found using a search engine and nine games from the POG data set (cf. Appendix B.1). After downloading and setting up the games as required, each game was analyzed for the topics described in the previous section. The summarized results are presented in Table 10.

Results show that only a few games include detailed explanations of conceptual knowledge. In addition, there are only two games (NoPhish and Anti-Phishing Phil) that focus on URLs and explain how to determine the registrable domain (RD) of a URL. Only NoPhish goes beyond this and takes apart the structure of URLs in detail, making it the game with the most detailed explanations. However, since NoPhish is only available in German, its reach may be limited. While Anti-Phishing Phil also includes detailed explanations on how to locate the RD, it, however, misses some details compared to NoPhish, e.g., information on subdomains and the separation of different domain labels.

For games in the *Email* category, none of the games conveyed any detailed conceptual knowledge, e.g., none of the games explain email headers or how to verify email authenticity. As for attachments, seven games include information on malware in attachments or links in emails. Two games also address other message types beyond email (e.g., instant messaging).

[19] https://github.com/, accessed on 07.07.2021
[20] https://codecanyon.net/item/anti-phishing-awareness-game/20935555, last accessed on 2020-04-16

Table 10: Analysis of game content ($n = 13$)

	Anti-Phishing Phil [She+07]	ATMSG/CSAG [Huy+17]	Birds Life [WJZ18]	codecanyon[20]	cyberaware [GKG15]	CyberCraft [Lu18]	GHOST [KW18]	NoPhish [BC14]	OpenDNS[21]	Sophos[22]	whatdothack [Wen+19]	whatthehack [Gey19]	WithGoogle[23]
Registrable domain	3	1	0	1	0	0	0	3	2	0	1	0	2
Deception	2	1	0	1	0	0	0	3	1	0	1	0	2
Other URL features	2	1	0	1	0	0	0	3	2	0	1	0	2
Email traits	0	1	2	0	0	1	0	0	0	1	2	0	2
Sender spoofing	0	1	2	0	0	0	0	0	0	1	1	0	1
Advanced attacks	2[a]	2[b]	0	0	0	0	0	0	1[c]	0	2[c]	0	0

[a] Hex IP-addresses, [b] Pop-ups, [c] Hosting service abuse

Hypothesis H5

Lastly, results show that games in the *Other/Various* category do not offer any detailed explanations about phishing or anti-phishing protection strategies. Overall, this leads to the rejection of **H5**, as it seems that most games are either used to motivate different types of educational material or to test knowledge that was acquired differently.

Hypothesis H6

An additional finding is the lack of advanced phishing techniques in available games, thus, supporting the acceptance of **H6**. None of the games include information on IDN or percent-encoding in URLs, and only a few include hosting platform abuse. This might be due to the low expected occurrence of this type of attack in phishing or due to the fact that the emergence of advanced techniques is a comparatively recent development. On the other hand, six games include examples of benign websites with uncommon characteristics, e.g., the use of 'www3' as subdomain instead of 'www' or information on unexpected domain names (e.g., 'dropboxmail.com' instead of 'dropbox.com'). Including these examples could potentially demonstrate to users that benign websites might also exhibit uncommon behavior, thus reducing false positives in their decisions.

Many games also include hints on additional protection strategies, like using a search engine to determine the authenticity of a website (3 games) or hovering over a link to display the actual destination (3 games). Using MFA was recommended less often, and only one game included it as a method to protect against phishing.

All in all, it should be noted that only the content of the games was analyzed, and it is not claimed that including or omitting these topics has a particular effect on learners. However, the relevance of the selected topics is still valid due to the body of knowledge on phishing.

5.3.5 Discussion

In this section, the results of the preceding three analyses will be discussed, and respectively conclusions are drawn regarding the before-mentioned research objective statements.

[21] https://www.opendns.com/phishing-quiz/, last accessed on 2020-04-16
[22] https://www.sophos.com/en-us/lp/games/play-spot-the-phish.aspx?cmp=35375, last accessed on 2020-04-16
[23] https://phishingquiz.withgoogle.com/, last accessed on 2020-04-16

Summary of results

The analysis of target groups and educational contexts revealed that many games are suitable for end-users in informal learning contexts, thus, leading us to accept **H1** and **H2**. It also showed that most of the games require specific categories of BRT, namely remembering and understanding conceptual and factual knowledge (**H3** accepted), regardless of their scope or digital nature. However, the majority of games do not require procedural or meta-cognitive knowledge, which is why **H4** is rejected. Furthermore, considering only digital games focusing on phishing, those require an even smaller range of activities (**H3a** & **H3b** rejected). Lastly, **H5** is rejected since the effect of game-based learning on motivation is not used to introduce more detailed content, and **H6** is accepted since the games lack information on advanced contemporary phishing techniques.

Limited availability of games

The main limitation of the findings is that only a few games are actually available and accessible. It severely hinders attempts to extract general findings. Thus, it hurts the attempts to replicate research and improve the development of anti-phishing learning games. Nevertheless, the hypotheses were tested and warrant some discussion and interpretation.

Target groups and educational contexts

As identified in Chapter 5.3.2, end-users and non-CS students are the most often addressed target groups of anti-phishing learning games in the POG data set. These target groups can be characterized as having no or little prior knowledge in IT security or CS while using IT systems and the Internet in their daily activities. This way, it is not surprising that research projects focus on educating these target groups to enable risk-aware behavior and be able to recognize and avoid phishing attacks.

Similarly, games are designed for employees in fields other than IT. On the contrary, only a few games are intended for CS students or IT employees, which are both target groups that usually have prior knowledge in IT security and CS. Here, it could be that anti-phishing education for target groups with prior education in CS or IT security is perceived as less complex, and thus, game-based learning might not be utilized as much. As presented in Chapter 5.2.3, the results show, however, that there are learning games that are intended for CS students and IT employees or professionals. Similar to the analysis of target groups, results show that most games are intended for informal learning contexts. This is immediately linked to the results of the target group analysis since end-users would play anti-phishing learning games in informal learning contexts. Regarding formal educational contexts, it is distinguished between different levels in school as well as college/university level and corporate contexts. The results show that only a few games are designed for schools. Meanwhile, more games are designed for college/university and corporate learning contexts. Depending on the coverage of CS education as well as the use of digital media in school, anti-phishing learning games might not have gotten a real chance yet. With an increase in CS education in schools, anti-phishing education might become more important for the target groups of students in school to enable risk-aware behavior and phishing protection strategies.

Learning content

The learning content of the accessible games is comparably easy to discuss, as it concerns verifiable technical information. The analysis results show that only a few games offered detailed explanations that would allow users to gain a fundamental understanding of key concepts. Thus, most games would need additional, accompanying educational materials to provide suitable anti-phishing education. In the analyzed games, the most detailed explanations seen while playing the games focus on URL classification, particularly locating the registrable domain. However, more advanced techniques are also not covered, although some games offer examples. Although email phishing was a topic of multiple games, none offered detailed knowledge of phishing emails. As a

result, this common attack vector is severely underrepresented. Evaluating the severity of these gaps in anti-phishing education poses the question of how much detail and depth would be sufficient to enable users to protect themselves from phishing attacks. This question is difficult to answer, as attack and defense knowledge co-evolve, changing the necessary knowledge over time. Also, it is hard to discern which contents are most efficient and effective at improving security, necessitating further research. However, it would be beneficial to teach sustainable knowledge and skills for users to rely on and enable users to adapt and learn more over a period of time.

Learning goals

As analyzed and described in Chapter 5.3.3, learning goals and game mechanics of the analyzed games are suitable to convey conceptual knowledge since learners need to remember and understand concepts within the game. Many games implement a binary decision mechanic in which learners have to classify a URL or email as either legitimate or phishing. However, only a few games actually use mechanics to convey conceptual and procedural knowledge in-depth and therefore may impede the creation of suitable mental models that help users develop secure, risk-aware online behavior. This is referred to as the conceptual-procedural principle, where the idea is to present knowledge (conceptual) and then use it to solve in-game problems (procedural). While mentioned in some publications, the principle is not often leveraged in available games. The results show that the procedures needed to advance in the games are generally not applicable to deal with real-world attacks, i.e., pointing at a phishing URL does not solve the problem in an actual phishing attack. Also, none of the games focusing on phishing made the learners create or evaluate knowledge. However, in reality, end-users need to create, apply and evaluate their own security procedures against phishing attacks. This poses the question of whether it is possible to design games that support these skills. Alternatively, it could be considered to embed available games into appropriate learning settings where additional resources, briefings, and debriefings support learners while at the same time conserving the games' motivational advantage.

Conclusion

In conclusion, the available anti-phishing learning games are limited in their information content and educational potential since they are missing detailed information on phishing threats and do not properly train secure, risk-aware behavior. While most games are designed for an informal learning context, the review results show that they fail to serve as independent learning environments and can only provide an assessment of prior learnings. Most games rely on conveying factual and conceptual knowledge, but they miss out on teaching procedural or meta-cognitive knowledge. Often, gameplay requires learners to simply decide whether a URL or email is considered phishing or not. The assessment capabilities are somewhat limited, as a binary decision does not give insights into the learners' misconceptions or correct application of knowledge. Lastly, it can be argued that a lack of detailed information might thus lead to confusion when presented with unknown or unexpected situations (e.g., an advanced attack using Internationalized Domain Names (IDN) or a link using an IP address). In contrast, inadequate understanding of secure behavior can lead to fatigue (e.g., if users check the URL after every click) or harm (e.g., failing to check for indicators of malicious websites after being redirected).

5.4 Conclusion and Identification of Further Research Gaps

This section concludes both systematic literature reviews on learning games for security education and anti-phishing education. The results of both reviews are summarized, and further research gaps are identified (in addition to those presented in Chapter 3.4).

Learning games for security education

First, the results of the systematic literature review of learning games for security education show that although many games are mentioned in the literature, the availability of the games is limited. As such, only 39 out of 99 games were available to play at the time of the review. The games were designed for various target groups and intended to be used in different educational contexts, e.g., high school students in formal learning contexts or end-users in informal learning contexts. Among the available games, only 20 games were suitable for end-users; however, a review of the games' content revealed further limitations. For some cases, the games' learning content was rather limited and often presented out of context. Depending on the game, either the games' duration or the intended target group (e.g., children) can be reasons why there is no detailed, extensive learning content. Meanwhile, some games are intended for target groups in formal educational contexts (e.g., students in college/university) that provide more detailed learning content, e.g., CyberCIEGE [ITA05]. However, due to the large variety of topics covered by the term security education, it is difficult to assess the content quality. Overall, the results of the first review were rather inconclusive, and a second review with a focus on anti-phishing learning games was proposed.

Learning games for anti-phishing education

The systematic review of anti-phishing learning games was done in a similar manner to the first review. However, the analysis was extended and included an analysis of learning goals and learning content. As such, the POG data set was generated through exhaustive search and filtering (see Appendix B.1). Based on this data set, the intended target groups and educational contexts were analyzed. Results show that the majority of anti-phishing learning games are designed for end-users with little to no prior knowledge in IT security, and they are most often intended to be used in informal learning contexts. Furthermore, the analysis of learning goals shows that most anti-phishing learning games convey factual and conceptual knowledge but do not focus on procedural knowledge. None of the games make learners create or evaluate knowledge. Similarly, many learning games only implement a binary decision mechanic in which learners are required to decide whether a URL or email is legitimate or phishing. Here, assessment is limited and does not give insights into the learners' decision processes. Regarding the games' learning content, different topics in anti-phishing education are covered, e.g., phishing attacks using URLs or emails. However, the games fail to convey detailed conceptual knowledge as they fail to offer detailed explanations to help users gain a fundamental understanding of key concepts. The most detailed explanations observed during gameplay focus on URL classification, particularly locating the registrable domain. More advanced techniques are not fully covered, although some games offer examples. In conclusion, the results show that although existing games are suitable for end-users in informal educational contexts, they are limited concerning their focus on factual and conceptual knowledge, and they fail to include detailed learning content and advanced techniques. Since many games rely on a binary decision mechanic, they do not offer insights into the learners' decision processes. To this end, the assessment of learned knowledge and skills is limited, and no problems or misconceptions can be identified.

Conclusion

Overall, the two reviews presented in this chapter indicate that although a significant amount of research was done in the past, the availability of games is limited and of those games that are available, learning content and gameplay present different shortcomings. While many of the available games are suitable for end-users with little to no prior knowledge in IT security, the covered learning content is often not detailed enough to support learners in gaining a fundamental understanding of how phishing attacks can be detected. Furthermore, existing games mainly focus on conveying factual and conceptual knowledge while addressing only lower-level cognitive processes according to

BRT. There is a lack of games exploring higher-level cognitive processes and teaching procedural knowledge. Lastly, the use of limited game mechanics prevents learning about learners' decision processes and possible misconceptions and problems. This results in an untapped potential for a more detailed assessment of learners' knowledge and skills. Either a redesign of existing games or a completely new design incorporating different game mechanics and identifying learning goals covering the complete BRT could provide advances in closing these research gaps. Further, personalization can be explored to enrich gameplay and the learning experience.

6 Design and Implementation of Learning Game Prototypes

This chapter covers the design and implementation of two new learning game prototypes as well as a third game prototype serving as a baseline implementation similar to related work. First, the design goals, the selected learning content, and respective learning goals of the new learning games are described. The implementations of the first two prototypes are presented, including the tutorials, levels, and details regarding the development. Next, results of a preliminary evaluation of the game prototypes are discussed. Lastly, the third game prototype is presented. The different learning game prototypes are later used for evaluation and comparison of different game mechanics (see Chapter 8) as well as for the adaptation of learning game content using the personalization framework (see Chapter 7).

Large portions of this chapter were previously published in [G] as well as [B]. The design and implementation of the different game prototypes was collaborative work with Vincent Drury, preceded by different student thesis projects.

6.1 Design

In this section, the design goals for the development of new anti-phishing learning games are derived from the results of the reviews presented in Chapter 5. Furthermore, the learning content is specified, and learning goals are formulated.

Summary of the state of the art

While there are different learning games designed for end-users in informal learning contexts and, therefore, suitable to the target group of this dissertation, the actual implementations of these learning games are often unavailable (as derived in Chapter 5). Thus, reviews were limited to related publications about the games. Only nine games within the POG data set were available for download or to play online (see Appendix B.1). Overall, existing anti-phishing learning games mainly address lower-level cognitive processes according to BRT, e.g., remembering or understanding factual and conceptual knowledge. In addition, the learning content of available games was analyzed by playing them, and the results revealed an insufficient level of detail regarding phishing attacks, their recognition, and potential countermeasures in most games. Furthermore, existing games lack information on advanced contemporary phishing techniques, which might leave players vulnerable when faced with novel types of phishing, which were not part of the games. Regarding the gameplay of existing games, the use of different game mechanics is limited since most games rely on a binary decision mechanic, where players are presented with a potential phishing URL or email and have to classify it as phishing or benign (i.e., a binary decision as there are only two options two choose from). To this end, existing games fail to offer sufficient practice and detailed assessment of players' decision processes. Since none of the games were available as open source implementations, it was not possible to extend existing games by improving game design and integrating personalization.

Design goals

Based on the findings of the review work explained above, the decision was made to development new learning game prototypes for anti-phishing education. As such, the following major design goals were posed:

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

Different game mechanics

In order to fulfill these two major design goals, two novel anti-phishing learning games were implemented. For the content of both games, the sub-topic of URL phishing was selected to provide comparability to existing anti-phishing learning games (see Chapter 6.1.1). The goal for the first game prototype was to extend the often used binary decision mechanic and provide multiple options to classify a given URL or even to discard it. The hypothesis is, that this would allow for a more accurate assessment of how players decide and when they are unable to decide. Similar to existing games, feedback is provided for correct and incorrect classifications. For the second game, the goal was to replace the currently favored game mechanic of classifying URLs by introducing a puzzle or combination game mechanic. As such, players have to create their own URL for a given task by combining URL pieces. The hypothesis is that this would allow for addressing higher cognitive-order processes. Feedback is given based on a set of automated checks that verify the correct application of different manipulation techniques.

Overall, the first game's design follows an analytical approach, where the decision scheme requires players to analyze a given URL and classify it based on given categories. A more constructive approach was chosen for the second game by changing the game mechanics and requiring players to create their own URLs by applying different manipulation techniques. 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 in the scope of a user study. If a combined prototype would be 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. It should be noted that design and implementation of the different game prototypes was approached rather practical without using an explicit game design framework. As such, this presents a limitation to this work and allows for redesign steps in future work. However, in order to implement personalization (see Chapter 7) and compare the different games (see Chapter 8), the development of prototypes was necessary.

6.1.1 Learning Content

For the learning content of both games, the topic of URL phishing was chosen to enable the development and design of games that are comparable with existing games, most of which cover different aspects of phishing URLs or emails. The main requirement for the selected learning content was to provide knowledge and skills that are robust to adversarial influence and possible for the end-user to understand while maintaining general applicability. URLs represent suitable learning content because they uniquely identify a website, thus prove robustness, and are also applicable in a wide range of potential attack scenarios. URLs can be used to analyze a potential phishing website, regardless of how the website is reached (e.g., link in email, text message or social media post). In addition, the knowledge about URLs can be used in several other

scenarios, e.g., when analyzing email headers to determine the sender or when inspecting certificates to identify their origin [DM19].

Figure 2: URL structure[24] with highlighted components: Subdomain, Registrable Domain (with its TLD) and Path

URL structure

The games aim to teach the structure of the URL[24] by focusing on three components: (1) subdomains, (2) registrable domain (RD), and (3) path (see Figure 2). Other URL parts could also be considered (e.g., authentication information or a port specification in the hostname) but were not included in the game prototypes, as they are less common, and in particular, did not appear on any legitimate login pages that were used as examples in the games. The example URLs are based on the login websites of a set of existing services that are popular in Germany (e.g, 'Ebay', 'Paypal' or 'Amazon'; based on the Alexa[25] and Tranco[26] lists). Each part of the URL structure is introduced together with illustrative phishing URLs, which can be detected by understanding the corresponding part of the URL. Therefore, both learning games teach how to identify the registrable domain of a URL, as it is the discerning factor to decide between legitimate and phishing URLs. More generally, both games focus on the required knowledge and skills of URL parsing (as defined in Chapter 2.1.3).

Manipulation techniques

To support a better understanding of how to identify potential phishing URLs, the games include a number of techniques attackers use to construct phishing URLs. The techniques are referred to as *manipulation techniques* as they are used to manipulate a benign URL and create a malicious URL, e.g., 'combo-squatting': 'ebay-service.com', 'addition':'paypall.com'. These manipulation techniques are based on real attacks that were found in the wild and described in related work (cf. [Rob+19; Agt+15; Kin+17; Oes+18]). In addition, they capture the structure of URLs as taught in the games, thus creating a foundation of what to look out for in phishing attacks while strengthening the players' understanding of URLs. In the scope of this work, it is assumed that attackers typically construct phishing URLs that make the user believe they lead to a specific benign target domain, e.g., a trusted bank or online shop where the user has an account. This target domain is referred to as the *original domain* (for a more comprehensive description of a typical phishing attack, see Chapter 2.1.3).

The manipulation techniques included in the games are:

- `Subdomain`: Including (parts of) the original domain in the subdomain of the URL
- `Path`: Using a random domain as well as (parts of) the original domain in the path of the URL
- `IP address`: Using an IP (Internet Protocol (address)) address as the host part of the URL and (parts of) the original domain in the path of the URL

[24] as defined in `https://url.spec.whatwg.org/`, online, accessed 11.09.2021
[25] `https://www.alexa.com/topsites/countries`, accessed on 25.06.2021
[26] `https://tranco-list.eu/`, accessed on 25.06.2021

- `RD`: Modifying (e.g., adding to, swapping characters of) the RD of the original domain of the URL
- `Top Level Domain (TLD)`: Using the original domain, except for changing the TLD
- `Random`: The given URL is completely random; there is no connection to the original domain or any other suspicious keyword

In conclusion, the learning content of both games covers the threat of URL phishing, and the games aim to impart the knowledge and skills necessary to detect phishing attacks using malicious URLs. However, the learning content selection presents a limitation to this work. Although the knowledge and skills for recognition and countermeasures of URL-based phishing attacks may help in protecting against phishing, to protect against an actual attack successfully, users also require the awareness to apply the conveyed knowledge and skills at the right moment. This may be a behavioral change in their daily activities. Even if the developed learning games might facilitate such a behavioral change, it is not the main design goal or focus of the current prototypes for this dissertation and might therefore be studied in more detail in future work. This may include exploring different topics of anti-phishing education or improving the games selection of content regarding when and how phishers lure potential victims into disclosing personal information.

6.1.2 Learning Goals

Bloom's Revised Taxonomy

Based on the learning content described in the previous section, this section presents the learning goals of both learning game prototypes. For the formulation of learning goals, BRT is used as it provides guidance and structure for the determination of learning goals as well as support for the design of learning interventions [Kra02]. As introduced in Chapter 5.3.3, BRT can also be used to analyze existing learning interventions and to understand what type of knowledge and cognitive processes are addressed by the intervention. Learning goals are structured into six different cognitive processes: (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.

The learning goals for the two learning game prototypes described in this section cover a wide range of cognitive processes and types of knowledge. Regarding the knowledge dimension, covered learning content ranges from terminology, concepts, and principles to subject-specific techniques and methods, i.e., URL parsing and manipulation. This way, learning goals and activities are able to impart factual, conceptual but also procedural knowledge. In their current state, neither of the two learning games impart meta-cognitive knowledge, i.e., knowledge and awareness of one's own cognition [Kra02]. Depending on the game prototype, only a subset of cognitive processes in BRT is covered according to the learning goals. Table 11 presents all learning goals as well as the matching to both games. The wording of the learning goals follows a fixed structure, always starting with the prefix sentence: "After playing the learning game, players should be able to…". With the focus on different types of knowledge and higher-order cognitive processes, the design of both games is in alignment with the second design goal.

Table 11: Learning goals including their mapping to both games. Learning goals are marked with x if they apply to the analysis game (A) or the creation game (C). This table was previously published in [G] and [B]

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

6.2 Implementation

Following the posed design goals, two new learning game prototypes were implemented with the beforementioned learning content and learning goals. 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 referred to as the *analysis game*[27]. It follows an analytical approach similar to existing games where players have to analyze a given URL. However, instead of only deciding whether a given URL is malicious or benign, players have to sort the URL into a particular category based on the applied

[27] https://gitlab.com/learntech-rwth/erbse/analysis-game, accessed on 07.09.2021

manipulation technique. Therefore, the analysis of URLs addresses the first design goal described in Chapter 6.1. A first prototype of the game was developed as part of a student thesis project [11]. For the scope of this dissertation, the prototype was restructured and partially re-implemented. Here, learning game content was changed to fit the design decision described in the previous section.

The second game is called 'A Phisher's Bag of Tricks' and is referred to as the *creation game*[28]. Addressing the second design goal, this game requires players to apply common manipulation techniques used in phishing attacks to create their own malicious URLs while keeping the URL structure intact. The current version of the game was implemented based on an adapted, first concept originally developed in a student thesis project [5].

Tutor characters and digital storytelling

The structure of both games follows an alternating tutorial-level scheme where, specific portions of the game's learning content are introduced in a tutorial. Then, after completing a tutorial, players continue to practice and acquire the 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. While in the analysis game, the tutor character is embodied by a Roman figure which introduces the URL structure and different manipulation techniques, 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 at a minimum, and thematic decisions originate from the first implementations of thesis projects. However, digital storytelling elements and themes could be changed easily in both games by providing different assets (e.g., images for background or characters). For the scope of this dissertation, game design, especially digital storytelling, was not the main focus but instead used to make the games possibly enjoyable and more comprehensible. Instead, the focus was on the design and implementation of different game mechanics to enlarge the state of the art and contribute a different approach to anti-phishing learning games. As such, the choices in digital storytelling present a limitation to this work and could be revisited in future work.

In the following, the tutorial and level design, including gameplay and feedback capabilities, are described. Later on, technical aspects of the underlying framework and the development are presented.

6.2.1 Tutorial Design and Content

The structure of both games utilizes the concept of tutorials, i.e., a practical introduction to respective learning content. The knowledge is impaired by the tutor character, that guides players through the game by explaining the game mechanics and teaching the required knowledge to progress through the following levels.

Tutorial design

Tutorials are designed as self-paced, stepwise sections, in which players can follow the introduced learning content but also can go back to previous steps if needed. The tutorials of both games include interactive elements, where players hover over a given URL and discover hidden information regarding different components of the URL. This mechanic may enable players' discovery and analysis actions as they have to actively search and inspect given URLs. For the analysis game, this fits the gameplay required in the levels following a tutorial section. Figures 5 and 7 show two different tutorial sections of the analysis game and creation game. Each tutorial consists of an example

[28] https://gitlab.com/learntech-rwth/erbse/creation-game, accessed on 07.09.2021

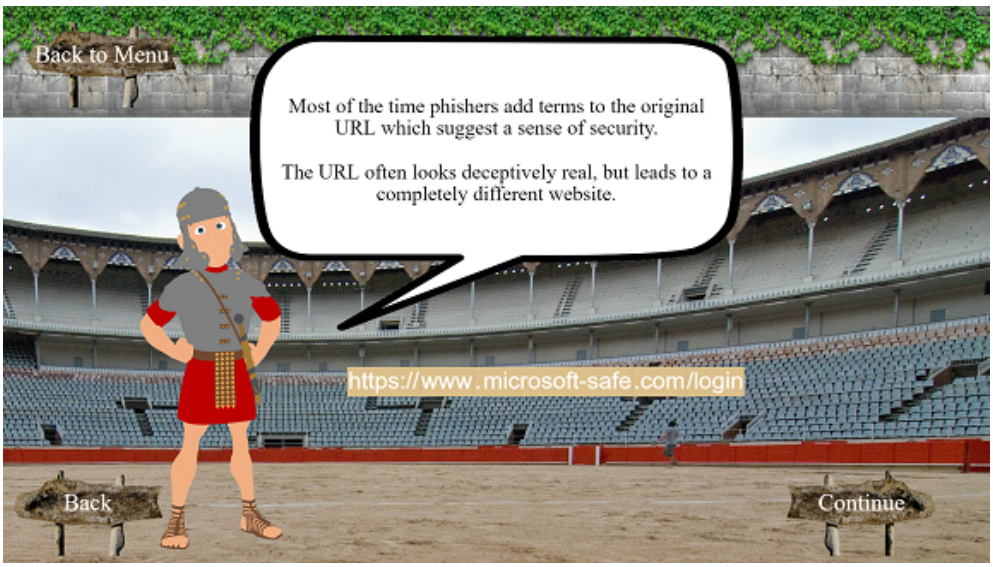

Figure 3: Tutorial (second level) of the analysis game. The tutorial text partially explains the manipulation techniques for registrable domains.

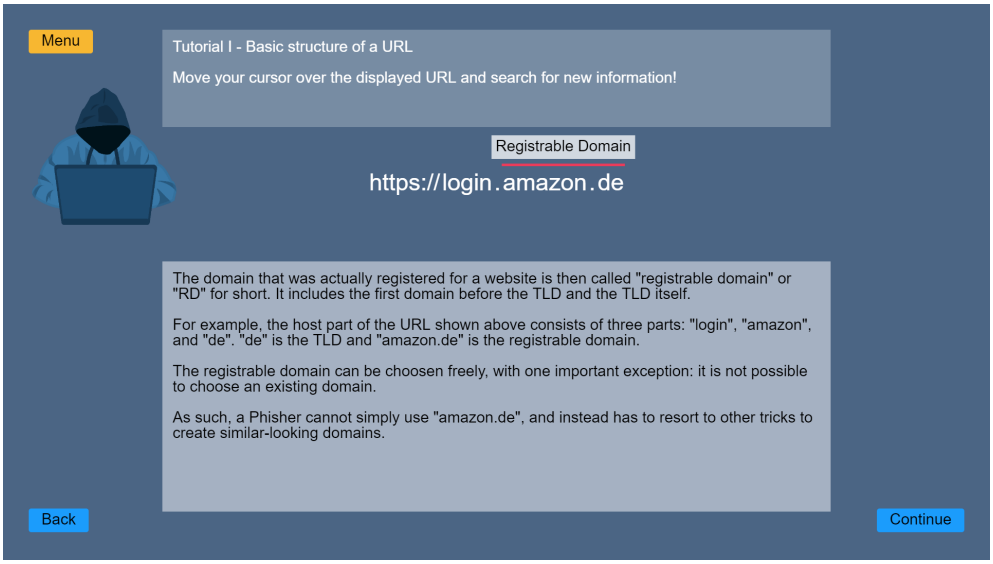

Figure 4: Tutorial (first level) of the creation game. The tutorial text partially explains the concept of registrable domains.

URL and a text explanation provided by the introduced tutor character. Furthermore, three buttons are arranged in the header and footer area of the screen: a button that takes players back to the game menu as well as two buttons to jump a step back or forward within the tutorial.

In the current game prototypes, the tutorials do not diverge into digital storytelling beyond the short introduction of the tutor characters. This way, tutorials are kept short so players would not get discouraged by lengthy explanations. Depending on the learning content, auditory explanations, as well as animated visualizations, could relax the amount of text presented in a tutorial.

Tutorial content

The analysis game consists of four tutorial sections covering the basic URL structure and key concepts like RD and IP addresses. Furthermore, the tutorials introduce different manipulation techniques for RD, subdomains, and paths. For the creation

game, the learning content is distributed using five tutorial sections. In the current implementation of both games, the tutorial content is similar but not equal. While manipulation techniques using IP addresses and random URLs are only included in the tutorial sections of the analysis game, they are not included in the creation game. This is because they offer an easy starting point for analyzing the URL structure when classifying different URLs and sorting them into different categories, but they do not add meaningful manipulations for players to perform in the creation game. Similarly, the manipulation of TLDs is an easy, first technique in the creation game, but it is often very complicated to classify without knowing the original URL, which is why they are not introduced in the analysis game. The differences in the tutorial sections of the games are later discussed in the evaluation of the game prototypes (see Chapter 8.1.4).

6.2.2 Level Design, Gameplay and Feedback

This section describes the gameplay of the analysis game and the creation game. Furthermore, feedback and assessment capabilities in both games are presented.

In both games, levels follow a tutorial section. They reinforce training and testing of the concepts and knowledge introduced in the tutorials. Therefore, levels offer a practice and assessment opportunity to players to apply the knowledge gained in the previous tutorial section and measure their performance. Similar to the tutorial design, the level design for both games divides the screen into three parts: the header area, the main content area, and the footer area. In the header area, the 'menu' button is provided in the upper left corner. While in the tutorials, the tutor character and tutorial content is presented in the main content area, and the footer contains buttons to jump back and forth, the level design for these areas differs in both games.

Analysis Game

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 5). In the main content area, the player is presented with a set of moving coins, which flip and reveal a URL (e.g., 'https://www.secure-paypal.com') and a context (e.g., PayPal) upon clicking on them. While the hidden URLs are represented by golden coins, the revealed URLs are placed on a parchment paper role, as they would not fit on the limited space of a coin. The set of URLs is randomized for each player and for each played level. This way, no set of URLs is the same, if players repeat a level. In order to classify a URL, players have to drag and drop the URL or coin into one of the buckets displayed in the footer area. Here, players can choose between a bucket for benign URLs and up to five buckets for different phishing URLs: `IP`, `random`, `subdomain`, `RD`, `path`. In addition, there is a bucket labeled 'no idea' which can be selected if players are not sure how to sort a given URL and want to discard it. This way, players are not forced to make a decision, if they are unable to classify it confidently and do not want to guess. The number of buckets for phishing URLs increases with each completed tutorial, and they correspond to the different manipulation techniques explained in Chapter 6.1.1. The benign and 'no-idea' buckets are present from the beginning and remain throughout the game. Right next to each bucket, a small question mark button is displayed. Upon click, the gameplay is interrupted, and a small text box (or 'info box') appears presenting a brief description of the URL category the bucket represents (see Figure 24 in the Appendix). This supports players who cannot remember the

difference between specific buckets and allows them to check before sorting incorrectly. After closing the text box, the gameplay continues.

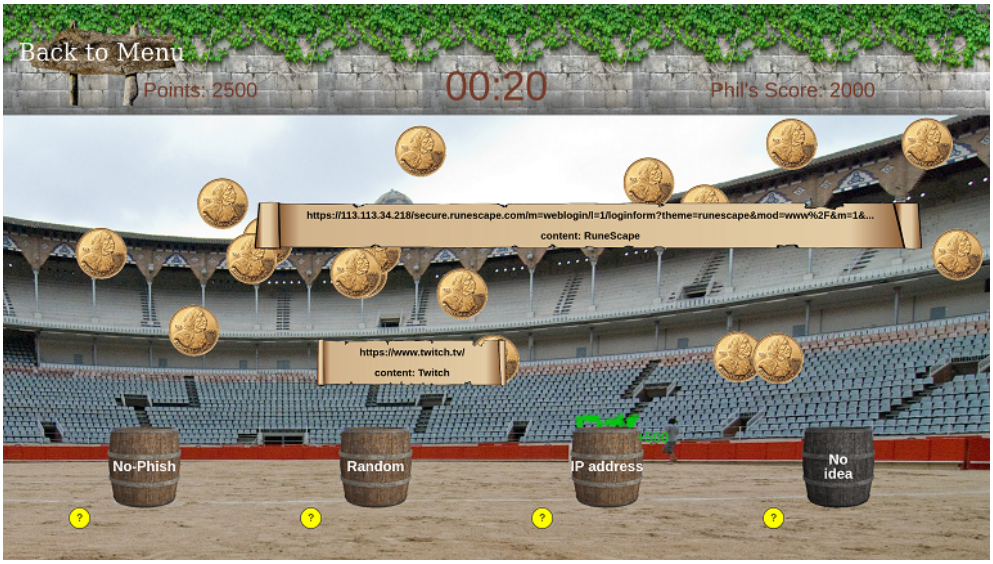

Figure 5: Screenshot of the first level in the analysis game where two coins are flipped showing the URLs and a green aura is shown above the bucket "IP address" indicating a recent correct classification by the player.

When dropping a coin into a bucket, players receive instant feedback in the form of a colored aura over the bucket:

- green aura for correctly classified URLs (see Figure 5)
- yellow aura if the classified URL was correctly classified as benign or malicious but not the correct manipulation technique (only correct tendency)
- red aura for incorrect classifications
- black aura, if players discarded the URL using the 'no idea' bucket

Furthermore, the aura also displays the increase of the players' score, i.e., plus points for correct decisions (500 points) or correct tendency (200 points), no points for incorrect decisions, or discarded URLs. 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 a coin is dropped into a bucket, a new coin appears randomly in the area above. For an overview, the header area contains the players' current score, the decreasing timer, and the preset score players have to reach. The values for points and timer present default values that can be changed in the game configuration. Besides Figure 5, more screenshots showing gameplay of the analysis game can be found in Appendix B.2.1.

When the time is up, the level terminates, and the game checks whether players achieved a higher score than the preset score. This serves as a challenge for players to classify as many URLs as possible within a given time limit. The default value in the current implementation is 2000 points, but it can easily be changed in the game's configuration. The game further checks that players did not make too many mistakes and achieved a high classification accuracy (default of 75 percent). This is implemented to prevent guessing. If neither the preset score nor the classification accuracy is met, players must repeat the level. If needed, they can go back to the tutorial as well.

If both conditions are met, players can advance and continue with the next tutorial-level section. To support players in the learning process, feedback is given after the level terminates. Feedback is presented for different types of mistakes (e.g., URLs were incorrectly classified as benign, or if URLs were classified as phishing but not in the correct type). See Figure 6 for a screenshot showing feedback in the analysis game (for additional screenshots, see Appendix B.2.1). The feedback screen supports learning from mistakes and tries to catch possible misconceptions to clear them up.

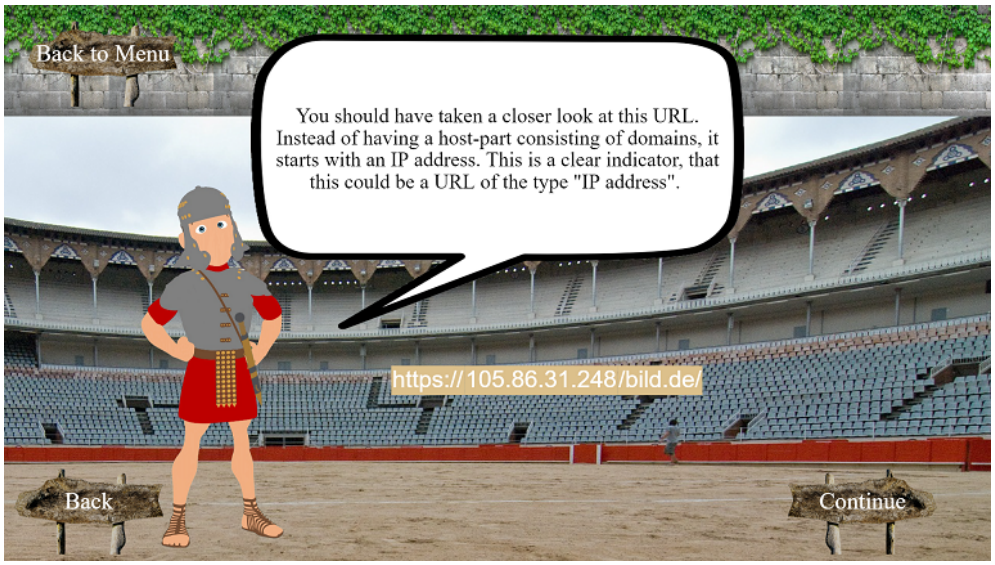

Figure 6: Screenshot of feedback in the analysis game.

For further assessment, all URL classifications and, as such, correct and incorrect decisions are logged and can be analyzed further to gain insights into the players' misconceptions (see Chapter 6.2.3 for more details on logging capabilities of the games). In the current implementation of the analysis game, log data is not analyzed in real-time during the game to provide further input for feedback or control the game flow. Here, personalization by adaptivity could be explored by analyzing players' in-game behavior using the log data and adapting the levels difficulty or overall path through the game (e.g., repeating a level if players fail to correctly classify URLs of a given type after learning about it in the previous tutorial).

Creation Game

A level in the creation game requires players to solve a set of tasks by creating their own URLs. A task is called *preset* and challenges players to create their own URL by combining different URL parts. In each preset, a task description provides relevant information on potential manipulation techniques or specific alterations that have to be used when creating a URL. The description is displayed within the header area of the level. 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 7).

The main content area of a level in the creation game contains a set of URL parts for players to use when creating their URLs. A URL part is represented by a character string of variable length. URL parts are depicted as small rectangular text boxes that can be moved via drag-and-drop. The set of URL parts contains more parts than required to solve the given preset. It contains common URL parts like 'https', '.com',

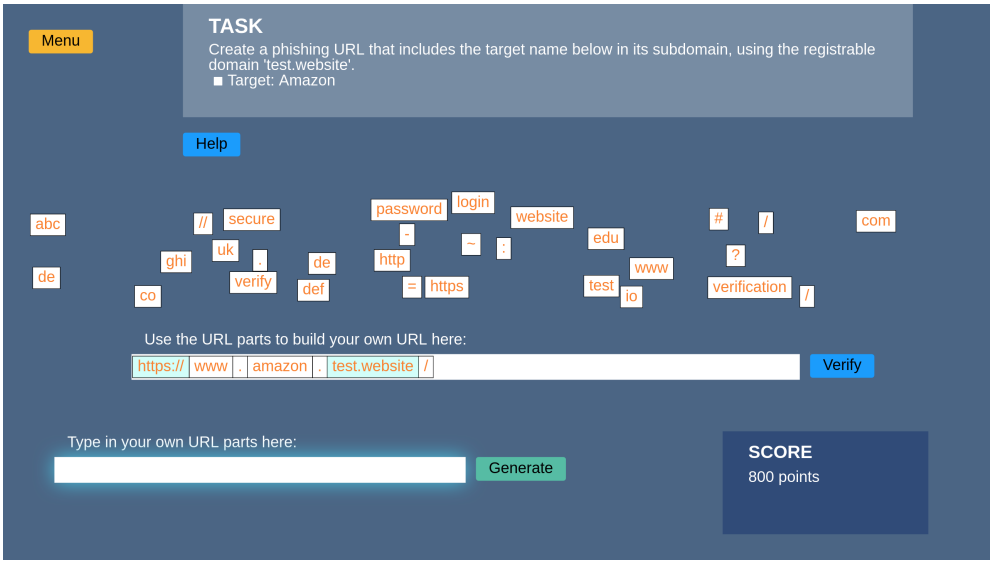

Figure 7: Screenshot of the third level in the creation game where multiple URL parts have been added to the URL bar including some own, custom generated URL parts (light blue URL parts).

'de', and 'http' but also a sufficient number of delimiter parts like '//', '.', or ':' to create valid URL fulfilling the preset. Further, the set contains arbitrary words to provide variability for players when combining them to create a URL (e.g. ,'secure', 'login', 'account'). While the content of the set is fixed for each level, the positioning of the different URL parts is randomized.

In order to create a valid URL and solve the task, players have to select different URL parts and drag them into a URL bar displayed in the lower part of the main content area. All URL parts within the URL bar can be sorted using a drag-and-drop action and shuffling parts to create the correct order. On the right-hand side of the URL bar, a 'Verify' button allows players to check their created URL and see whether it is a valid solution to the posed task. The footer area of the level consists of a status box displaying the players' current score. When advancing in the game, players can also access a text input box to create their own URL parts using keyboard input. This allows players to generate their own, customized URL parts as well as complete URLs as one part and move them to the URL bar to solve the task. As such, players can get creative in creating their own deceptive URLs. Existing URL parts and customized, generated URL parts can be used in combination to solve the presented task. The text input box and its 'Generate' button are displayed on the left-hand side of the footer area.

Similar to the analysis game, the creation game also provides help functionality. Just below the task description, a 'Help' button allows players to display the interactive URL structure used in the tutorial sections. Players can hover over the URL structure and see a highlighting of different URL parts. For post-hoc gameplay analysis, logging is also implemented for the creation game. As such, all player actions are logged and can be analyzed to understand player strategies and possible misconceptions.

Compared to existing games and the previously described analysis game, the creation game presents a different game mechanic in which players have to combine different URL parts to form a valid URL. This way, players have to actively create potential phishing URLs by applying manipulation techniques presented in the tutorial sections. When

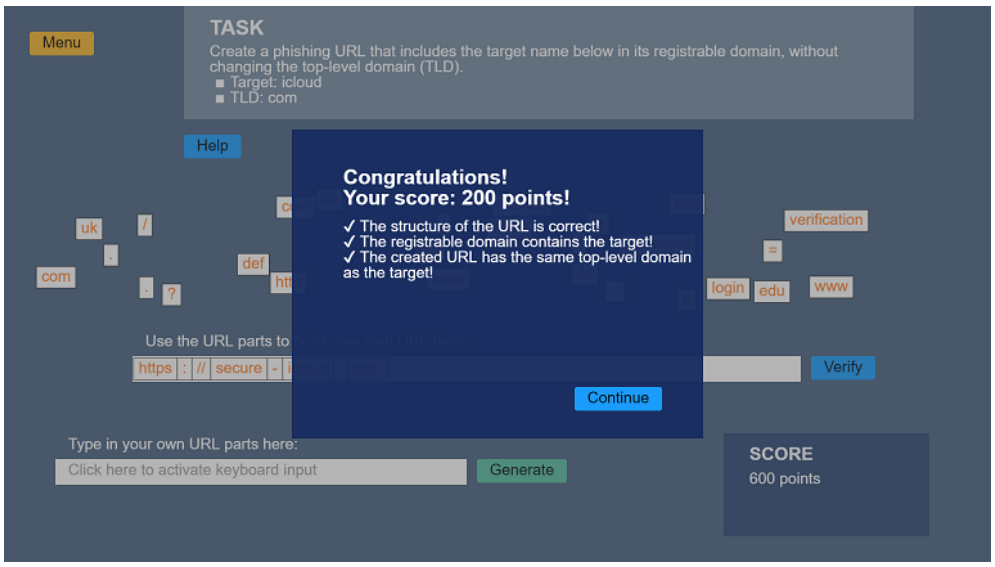

Figure 8: Screenshot of feedback in the creation game.

a solution is submitted, the game performs a set of automated checks on the created URL and provides feedback on which parts of the task were performed correctly and what has to be changed in case of a failed check. Based on the task, different checks are performed, and players receive feedback to solve the task in their next attempt. The checks include validity checks to ensure that the submitted solution is a valid URL and task-specific checks (e.g., examining whether a target was included in the RD or whether the created URL has the same TLD as the target). See Figure 8 for a screenshot of feedback in the creation game (further screenshots can be found in the Appendix B.2.2). Players receive points for passing different types of checks, indicating which tasks are more complicated and opening up the possibility to include non-mandatory checks that give additional points but do not have to be completed to advance in the game. Unlike 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 in which the game is used, a timer can be activated in the game's configuration.

6.2.3 Development

MTLG framework

Both game prototypes were developed using the MTLG framework (introduced in Chapter 2.2.4). Originally developed for games on multi-touch tabletop displays, MTLG is a modular and versatile game development framework for any browser-capable device, supporting any screen resolution. Games are deployed using a web server and, when accessed, immediately downloaded to the player's web browser. They run entirely on the client-side within the browser environment, and thus, no additional server capacity is needed after the initial download. Furthermore, MTLG provides various modules and extensions to enrich gameplay or enable learning analytics and feedback.

Logging capabilities

For evaluation and learning analytics, both games implement basic event logging functionality to capture all player actions and results achieved in the games. Log data is transmitted asynchronously to an external log server for further processing and analysis. To identify players, the log server generates human-readable and unique player IDs (i.e., pseudonyms). While this allows matching game sessions to external evaluation instruments (e.g., questionnaires), logging requires a stable Internet connection even though the games are running entirely in the players' web browser. If the Internet

connection is interrupted, log statements cannot be sent to the log server and thus, the log data will be incomplete. For future work, a service worker could be implemented to collect log statements that could not be sent to the log server successfully, in order to send them at a later point when the Internet connection is stable again.

In the first implementation, logging was implemented using a game-specific data format. However, the support of Experience API[29] (xAPI) for the MTLG framework enabled a switch to store log data in a standardized data format in a learning record store (e.g., Learning Locker[30]). For unique pseudonyms within the learning record store, a custom pseudonymity provider[31] is used. However, since logging and player IDs are only needed for evaluation purposes in future user studies, they can be deactivated for production environments.

6.3 Preliminary Evaluation

This section presents the results of a preliminary evaluation of both the analysis game and the creation game. The evaluation was performed in a minor development iteration and served as functionality testing to find potential issues and bugs. Similar to beta testing, a user group played the games, and feedback was collected. In December 2019, a group of 40 CS students ($N_{\text{Creation}} = 22$, $N_{\text{Analysis}} = 18$) played the games and answered a short survey about the functionality and user experience. Feedback from the evaluation was used as input to improve the implementation of the games and resolve presented issues and bugs.

Recruiting of participants

The participants sample consisted of a group of international students visiting RWTH Aachen University as part of an exchange program organized by the research training group UnRAVeL[32]. At the time, students were all enrolled in a CS Bachelor program at the university 'École normale supérieure Paris-Saclay' in Cachan, France[33]. While due to their enrollment in a CS study program, prior knowledge in CS was expected, the focus of the evaluation was primarily on functionality testing and finding potential issues and bugs. Thus, the suitability of the learning content was not evaluated.

Study setup

For the evaluation, the organizers of the exchange program divided the students into two groups of roughly the same size. While this was no true random assignment, it was ensured that the instructors of the study had no influence on how the group was divided. The study was conducted in two one-hour sessions, in which each group played one of the games, while the other group participated in some other unrelated activity as part of the exchange program. After one hour, the groups switched without any interaction between the participants. The study sessions were hosted at the InfoSphere student laboratory[34] at RWTH Aachen University, which was prepared with 24 laptops, one for each participant. In each session, participants were greeted by the instructors and introduced to one of the two game prototypes. The participants were instructed to play the game and to fill out a questionnaire after finishing the game. Afterwards, participants were asked to evaluate the game in groups of two and note down more feedback on different aspects of the games (see Figure 35 in the Appendix for detailed

[29] https://xapi.com/, accessed on 23.06.2021
[30] https://learninglocker.net/, accessed on 26.10.2021
[31] Hosted by the Learning Technologies Research Group at RWTH Aachen University; https://mtlg.elearn.rwth-aachen.de/pseudo/, accessed on 26.10.2021
[32] https://www.unravel.rwth-aachen.de/cms/~ofgh/UnRAVeL/, accessed on 03.02.2022
[33] https://ens-paris-saclay.fr/en, accessed on 03.02.2022
[34] https://schuelerlabor.informatik.rwth-aachen.de/, accessed on 03.02.2022

instructions given as a handout). 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 (see Table 34 in the Appendix).

Positive aspects of the games

Among the results, 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. For both games, participants praised the fun and engaging way to learn about URLs and phishing, the quality of the examples and detailed explanations. Participants also emphasized the importance of the topic and that the games could be useful for learners without prior knowledge (e.g., one participant noted that the games were "interesting, [and] I learned a lot about URLs, something I usually don't think about", while other participants noted "important lessons", "I think it is a good game for people who do not know about phishing" and "Good if no previous knowledge about phishing"). One participant described the experience of playing the creation game as a fun and engaging way to learn about phishing, while another participant highlighted how the combination mechanic invited tinkering with the URL manipulations.

Negative aspects, bugs and problems of the games

The participants also identified some bugs and issues during gameplay, 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 for the presets, and problems on slower devices that made it complicated to open coins to reveal URLs in the analysis game. As such, participants recommended replacing the coins by buttons or displaying the URLs immediately and without clicking on the coins, if the problems could not be resolved. Four participants criticized the explanation texts, which were received as too long and boring. Further, it was noted twice that the games are missing context information and advanced URL manipulation techniques. This might be due to prior knowledge, since the participants were CS students attending university. As such, they might have prior knowledge and were easily bored by the foundations of the URL structure and manipulation techniques.

Improvements and suggestions

The critique was also reflected in the suggested improvements, where some participants proposed interactive elements in the tutorials to avoid lengthy, boring explanations. Other proposed improvements include a competitive multiplayer mode and more variety in the use of game mechanics. Furthermore, three participants recommended including more URL categories and other topics of anti-phishing education (e.g., how to identify phishing emails).

Problems with unknown services

In addition to reported issues while playing the games, three participants also noted that they did not know all of the services that were presented in the analysis game. Due to possible difficulties in classifying URLs of unknown services, one participant explicitly asked for a list of legitimate domains to which they could compare the given URLs. This indicates potential problems in classifying URLs of services users are unfamiliar with. In the scope of this work, it motivates the personalization of anti-phishing learning games and to study whether personalized learning game content could provide a more suitable context in which players can learn about different manipulation techniques and can decide confidently between malicious and benign URLs.

Bug fixing in next development iteration

All reported bugs and issues have been resolved in the subsequent development iteration, and the initial tutorial design was improved by an interactive element, where players can hover over a presented URL and analyze its structure. Further improvements included a fix of the task generation and minor fixes to make the gameplay more fluent and with the intention to improve the user experience overall. While the selected participant group for this preliminary evaluation may not be the most suitable target

group for both games, their feedback was well received and valuable to improve the implementation and prepare the prototypes for a more extensive user study. The resulting implementations of both games were then used in a more comprehensive user study (see Chapter 8).

6.4 Baseline Implementation for Evaluation

This section covers a third learning game prototype, which serves as a baseline implementation when comparing different game prototypes. This game prototype presents a baseline with respect to existing anti-phishing learning games which implement a binary decision mechanic (as described in Chapter 5).

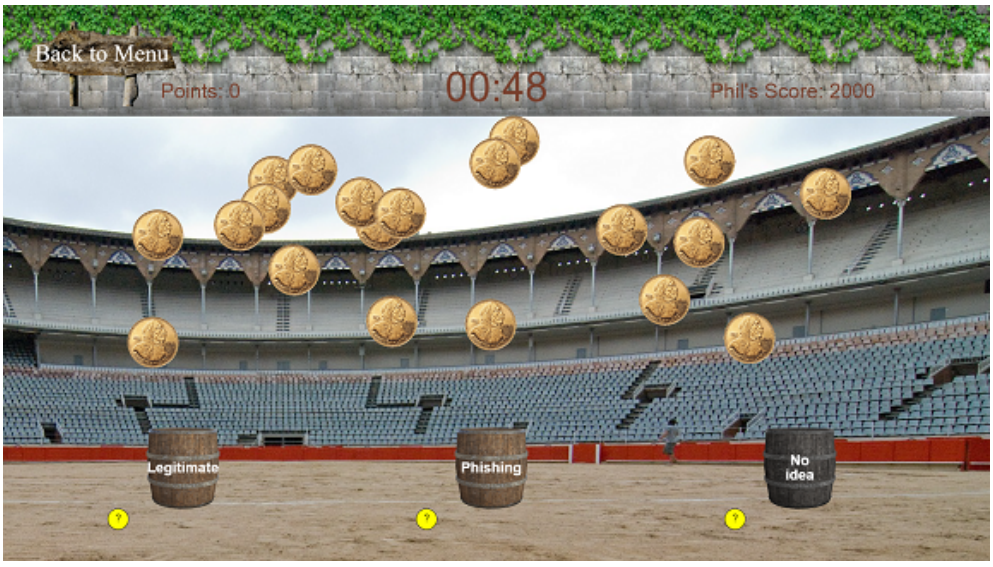

Figure 9: Screenshot of the first level in the decision game.

Decision game

The third learning game prototype is an adapted version of the analysis game and is referred to as *decision game*[35]. Figure 9 shows a screenshot of the first level in the decision game. Similar to both the analysis and creation game, the structure consists of alternating tutorial and level sections in which players first learn about a manipulation technique and the result type of phishing URLs and afterward are challenged in the levels to apply their gained knowledge. The tutorial design and the covered learning content remain the same as in the analysis game. While the level design is also very similar, the classification mechanic used in the analysis game is reduced back to a binary decision scheme which can be found in various existing games (e.g., [She+07; BC14; ALM15]). Therefore, only two buckets, one for benign and one for malicious URLs are available. To still allow players to discard URLs they are unable to classify confidently, a third bucket is provided (similar to the analysis game). Although this might extend the decision to be made between three choices, it can be argued that the decision mechanic is 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 (as described in Chapter 6.2.2). When a URL is classified as either benign or malicious, a colored aura indicates the classification result, and the players' scores increase if the classification

[35] https://gitlab.com/learntech-rwth/erbse/decision-game, accessed on 07.09.2021

was correct. After the given time is up, players receive feedback on their decision and, in particular, their mistakes. If the conditions for completing the level are met, players can proceed to the following tutorial. Otherwise, they have to repeat the level similar to the analysis game.

Since the decision game presents an adapted version of the analysis game in which players only have to decide whether a given URL is malicious or benign, it provides a comparable baseline implementation to the current state of the art, i.e., games relying on binary decision mechanics in which players only decide between phishing or benign. Since no existing anti-phishing learning game (with a focus on phishing URLs) was able to be adapted and aligned with the design and learning content of the analysis game and creation game, the decision game may serve as a baseline implementation that is similar to related work. While it is not the exact same as one of the existing anti-phishing learning (as identified in Chapter 5.3), it is equal to the analysis game and as such may allow comparing different game mechanics and evaluating whether there are differences in players' performances when classifying URLs. In the scope of this work, all three game prototypes have been studied in a comprehensive user study presented in Chapter 8.

7 Concept and Implementation of a Personalization Pipeline for Anti-Phishing Learning Games

This chapter covers the concept of a personalization pipeline for anti-phishing learning games and presents a reference implementation to be integrated into the games presented in Chapter 6. First, a conceptual description of the stepwise pipeline for personalization of anti-phishing learning games is presented. Next, the implementation of the personalization pipeline concept is described, including the overall architecture and its components. Lastly, its integration into the learning game prototypes is described, and possibilities for future extension and reuse for different learning games are suggested. The content of this chapter was previously published in [F] and [H].

7.1 Concept

This section presents the concept of a personalization pipeline for anti-phishing learning games. First, the motivation for personalization of anti-phishing learning games is briefly addressed, and then, the three-step personalization pipeline is introduced. The concept was previously published in [F].

Existing games and a lack of personalization

Current anti-phishing learning games that were found by the author of this dissertation do not support any kind of personalization as they do not account for learner characteristics. Literature reviews in the scope of this dissertation indicate that relevance and contextual are often not considered in game-based security education solutions, and anti-phishing learning games are limited in their design (see Chapter 5). A lack of personalization in anti-phishing learning games can lead to several potential shortcomings. Since most games rely on a binary decision mechanic where learners have to classify phishing URLs or emails as either benign or malicious, one potential problem may occur when learners do not know a given service behind the URL. As they do not have any point of reference to base their decision on, learners are unable to decide whether a given URL or email is benign or not. This might lead to frustration, hamper motivation, and could even negatively affect the learners' learning experiences. Also, learners might not feel more aware of phishing as a threat in their actual day-to-day activities. To this end, the games may fail to connect the presented tasks to the learners' phishing-relevant activities beside the game, i.e., the game content does not match the learners' online activity. The lack of personalization might, therefore, impair the playing experience as well as the effectiveness of learning material. In addition, phishing techniques are getting even more sophisticated. Attackers can use compromised accounts to send phishing emails from known senders and use information from data breaches to improve targeting while retaining high scalability. This results in more personalized phishing attacks, which might be even harder to prevent.

Research gap

With the widening gap between the personalization effort of attackers and defenders (e.g., educational interventions), the question arises of how personalization can be implemented in game-based anti-phishing education. To address this question, the concept of a personalization pipeline is proposed, which can be added to existing learning games and allows for the adaptation of learning game content to be more relevant to learners. The pipeline consists of three steps and includes (a) data collection from learners' devices, which can be combined with manually selected services in a personalization interface, (b) content generation to create personalized learning game content, and the final (c) content integration into anti-phishing learning games. Figure 10 depicts the concept of the personalization pipeline.

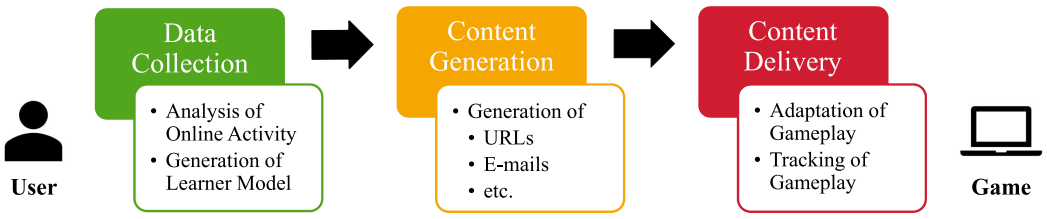

Figure 10: Concept of the personalization pipeline for anti-phishing learning games.

7.1.1 Data Collection and Learner Modeling

Objectives of data collection

The first step in the personalization pipeline is data collection. The objective is to collect informative data about the users' online activities (e.g., which services they use and what websites they visit). This information can then be used to generate a learner model or profile, which serves as an input to the pipeline's next step.

Data collection requires data and information gathering, which can be done using the following three approaches: (1) manual, (2) automated, or (3) hybrid data collection. Data sources can vary based on the applied approach.

Manual data collection

For manual data collection, one method could be asking the user directly, e.g., using a survey. This can lead to subjective but immediate insights. The targeted data source is the user's memory of past online activity (e.g., website visits, use of online services). By asking the user, only information about what the user remembers or wants to disclose can be collected. This leads to incomplete data about the users' online activity. Furthermore, data could contain untruthful answers, i.e., information on services or websites the user does not use or visit. In the end, manual data collection gives the user complete control over what data is collected, and the approach might not be perceived as invasive as automated or hybrid approaches.

Automated data collection

Automated data collection refers to the automated retrieval of user information from more objective data sources, e.g., the browser or email client. Possible data sources include the user's browser history, bookmarks, stored login credentials, and cookies. Beyond the browser, the operating system (either desktop or mobile) could provide information about installed applications and software (e.g., a banking software or music streaming service). The approach of automated data collection is more invasive and gives the user less control over what is collected and how. Although, the collected data might be accurate and complete (if users grant total access), problems can occur if multiple users use a device. If a browser is used by multiple users, visited websites and used services do not reflect a single user's online activity, e.g., if a couple or household share

a laptop and different people visit different shopping websites, payment services, and e-mail clients, automated data collection cannot match websites to users. Depending on the collected data, automated recognition of different users and possible matching of collected data to users could be implemented. This, however, goes beyond the conceptual work at this point and presents challenges outside the scope of this work.

Hybrid data collection

A combination of automated and manual data collection leads to a hybrid approach. While automated data collection is somewhat invasive and leads to unconfirmed findings (i.e., because multiple users could use one device), the hybrid approach allows for a validation step using manual data collection methods (e.g., asking the user to correct and confirm collected information). For hybrid data collection, first automated data collection methods are used to retrieve data from, e.g., the browser history. In a second step, identified services and websites are presented to the user for confirmation or to discard them. This way, users can iterate over the results of automated data collection and can manually refine them, which increases accuracy and trust. On the one hand, this introduces the subjective and possible dishonest user to the data collection process. On the other hand, it helps compensate for the shortcomings of automated data collection.

It should be noted that for all data collection methods, the users' consent is required first. Especially in automated and hybrid approaches, consent is not only required by the users who want to play a personalized game but also by all users of this device. This limits the usability of automated and hybrid data collection in large-scale field studies with users using their own devices.

Learner modeling

When data about the users' online activity is collected, a learner model can be created. As introduced in Chapter 2.3.3, a learner model refers to an abstract, structured representation of a learner, i.e., a player in the context of anti-phishing learning games. In the scope of this work, where collected user data contains information about the services and websites the users use or visit, the learner model can store this information as initial knowledge of the learner. Depending on what else can be collected, additional information could be added to the learner model, e.g., if manual or hybrid data collection is used, learners could be asked how familiar they are with a given service (e.g., they know and use the service, know and not use the service or do not know the service). The information about learners' familiarity with services could then be added to the learner model as well. Later, gameplay data (e.g., results from levels, error rates) can be added to the learner model such that personalization through adaptivity can take the updated learner model to adapt future gameplay. However, for the concept of this personalization pipeline, the backpropagation of gameplay data to enrich the learner model is not further outlined.

All in all, the result of the pipeline's data collection step provides input to the next step of content generation. The generated learner model reflects users' online activity as the learners' knowledge can be used for custom content generation to personalize the content of the respective learning game.

7.1.2 Content Generation

In the second step of the personalization pipeline, the learner model generated by data collection is used as input to different, game-specific content generators. This step in the pipeline aims to generate personalized game content that can be integrated into the respective games.

Depending on the game, the content type may differ. In the context of anti-phishing learning games, possible content types include URLs and emails. More complex content types that can be considered are phishing websites or complete scenarios, including digital storytelling. For the games described in Chapter 6, content generation may focus on malicious and benign URLs.

URL and email generation

When generating URLs or emails, an appropriate starting point is a benign service or class of services used to automatically creates different malicious or benign URLs or emails. For URLs, this can include common transformation rules used by attackers. As described in Chapter 6.1.1, manipulation techniques may include manipulation of the subdomain, path, registrable, and other URL parts. Transformation rules refer to string operations like swapping characters, concatenating, or inserting characters (cf. [Rob+19; Agt+15; Kin+17; Oes+18]). Emails are possibly harder to generate, as it requires natural language generation techniques. Instead, email templates could be used that are tailored to different service categories (e.g., financial services, retail/shopping). Templates can then be filled with information from the learner model. To this end, the learners' full name could be added to the learner model, in order to fill in email greetings (e.g., 'Dear John Doe' or 'Dear Mr. Doe' for learners named John Doe). Since emails often include phishing URLs in the email body, URL generation can be utilized during email generation; when customized, malicious URLs will be added to fill out the template.

Generation of complex content types

For the generation of more complex content types, approaches might be more elaborate and could be divided into multiple generation steps. Similar to embedding malicious URLs in email templates, scenario generation may include the generation of email communication and distinct elements for digital storytelling. The generation of phishing websites can be based on benign websites, e.g., by utilizing screenshots or source code.

Overall, content generation is a game-specific step in the personalization pipeline, and therefore, custom generators have to be provided to generate suitable learning game content. Finally, generated content is input to the next step of the personalization pipeline. In this dissertation, content generation will focus on the generation of malicious and benign URLs based on services stored in the learner model. Depending on applied transformation rules, one service can lead to large sets of URLs.

7.1.3 Content Delivery

The final step of the personalization pipeline is content delivery. The input to this step is generated learning game content from the previous pipeline step. This step aims to integrate generated content into the respective learning games. Depending on the game, this might include an adaptation of the gameplay.

Level generation and adaptation of gameplay

In order to embed generated content into any game, suitable interfaces need to be provided. One method of integrating generated content (e.g., URLs) into gameplay is generating custom levels. A level generator can utilize the game's configuration as well as the current game state (i.e., the learner's current state within the complete game) and combine it with generated content to create the next level. This can include a cyclic relation in which, after completing a level, the level generator provides the next level depending on the game state and generated content. Here, personalization through adaptivity can be considered since dynamic level generation allows for minor incremental adaptations, e.g., to keep learners in a flow state or divert game paths based on the

learners' decisions in the previous level. For the scope of this dissertation, personalization is achieved by replacing default game content with personalized, generated content based on the learner model.

With the integration of generated learning game content and possible adaptation of gameplay, personalization is achieved. Meanwhile, to evaluate the concept of the personalization pipeline and the effects of personalized games for anti-phishing education, the implementation of the personalization pipeline is required.

7.2 Implementation

In this section, the implementation of the concept of a personalization pipeline is presented. The three distinct steps of the pipeline are outlined by describing the implemented modules and components. Lastly, the interfaces and the data flow between different components are described to explain how personalized learning game content is generated and made available for the games. Details on the modular architecture of the personalization pipeline were previously published in [H].

Modular, component-based design

The implementation of the personalization pipeline follows a modular, component-based design where each stage of the pipeline is represented by one module consisting of multiple components. Following the single-responsibility principle, each component serves a specific task or purpose. Data flow is managed using simple interfaces between components. Figure 11 depicts a component diagram of the architecture, as well as the data flow between different components or modules.

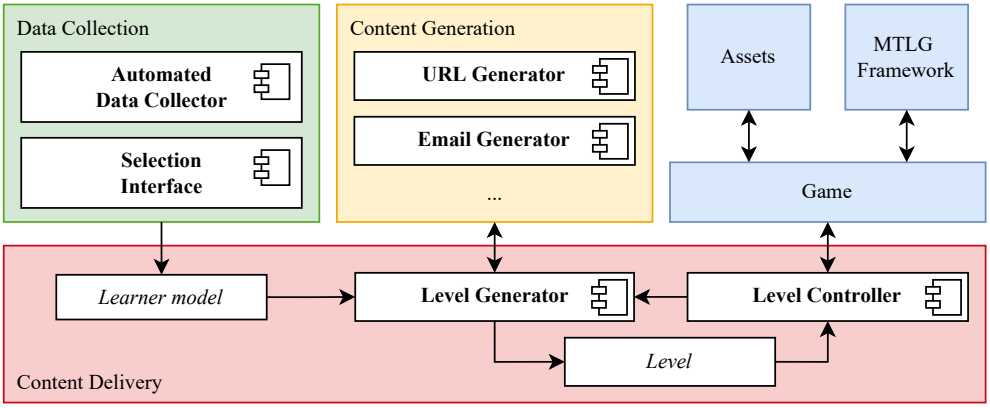

Figure 11: Component diagram of the proposed architecture. Modules of the pipeline include multiple components and are depicted using different colors. Arrows represent possible data flow between modules and components.

In the following, each module with its different components is presented. Further, data flow is explained to understand how the personalization pipeline is realized, and personalized learning game content is embedded into the learning game prototypes described in Chapter 6.

7.2.1 Module for Data Collection

All personalization efforts begin with creating a learner model, which is implemented in the data collection module. It does not require any specific input (apart from configuration parameters). Instead, it creates a learner model that contains relevant information

about the learner (e.g., the familiarity with services) and provides useful functions to access different learner characteristics (e.g., a list of known or unknown services).

As described in Chapter 7.1.1, three different approaches for data collection can be distinguished, i.e., manual, automated, and hybrid data collection. In the current implementation of the personalization pipeline, two components are provided in the data collection module: (1) a *Selection Interface* for manual data collection and (2) an *Automated Data Collector*. Both collect relevant information for the creation of the learner model.

Selection Interface

In order to support manual data collection, as one of the methods of the data collection module, the current implementation of the personalization interface contains a *Selection Interface* component. A first prototype of the selection interface was designed in a student thesis project [7]. For the context of this work, the design was adapted, and the prototype was reimplemented to fit in the architecture of the personalization pipeline.

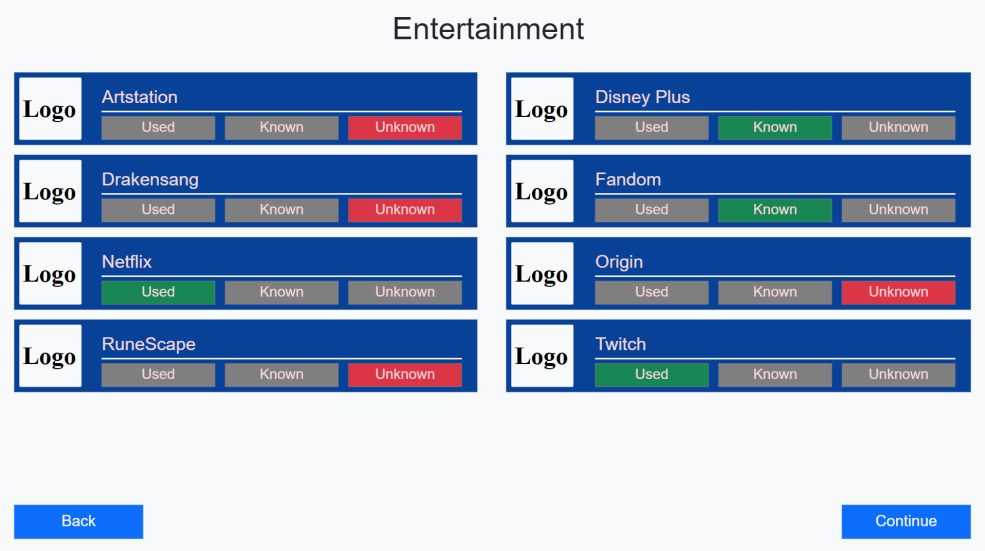

Figure 12: Selection Interface showing the category 'Entertainment' with selected options and placeholder for service logos.

Manual selection

This component provides a customizable selection interface showing a set of predefined services (e.g., 'Netflix', 'eBay', 'Twitter') of different categories (e.g., entertainment, shopping, communication). As shown in Figure 12, services are grouped in categories and displayed using pagination (see Table 35 for a complete list of all categories and services provided in the selection interface). For each service, the name of the service and a respective logo is displayed (if the service offers a logo). This way, learners can easily click through different categories and make their selections. Learners can select different options depending on their familiarity with a respective service. As such, learners can either select the 'used' option if they use a given service, 'known' if they know it but do not use the service, or the 'unknown' option if they do not know the given service. Through this selection process, the learners' familiarity with different services can be determined. As a result, the selection interface provides three lists of services with different levels of familiarity. Depending on the configuration of the selection interface, an option needs to be selected for either all services or learners can only select services for which they want to share their familiarity. However, if no option

was selected for a service, there is no knowledge about the learners' familiarity with this service, and it needs to be discarded before starting the learner modeling process. In the case that learners' do not use all three different options when completing the selection interface, one or more lists of service familiarity can be empty since no service was selected with the respective familiarity level, e.g., a learner did know or use all services; thus, the list of unknown services is empty.

The learners' familiarity with different services serves as input to the learner modeling process. Using the selection interface component, the data collection module supports manual, deductive learner modeling based on the learners' familiarity with different services.

Automated Data Collector

In addition to the manual approach, the data collection module also provides an automated, inductive approach for learner modeling using the *Automated Data Collector* component. The component is realized through a browser extension that needs to be installed to the learners' personal browser. The installed browser extension has immediate access to the learners' browser history and can extract all visited websites over a given period of time. Beyond the browser history, the extension can access learners' bookmarks and downloads. For security reasons, browser extensions cannot access stored account credentials (i.e., because otherwise, malicious code could extract stored account credentials and make them available for adversarial actions). The first version of the browser extension was developed in a student thesis project [6]. Afterwards, the implementation was integrated into the architecture of the personalization pipeline.

Analysis of browser history

Compared to the manual approach using the selection interface, the automated data collection is limited to the information of website visits. Based on visited websites, the browser extension can extract a set of services.The services are identified by extracting the hostname with relevant subdomains (except 'www'). As such, services like 'calendar.google.com' and 'mail.google.com' are identified as two separate services instead of generalizing it to 'google.com'. This results in a more detailed set of services than compared to the services pre-configured in the selection interface. The identified services can be categorized as services learners use or know. In an advanced analysis step, the browser extension can also analyze the frequency of website visits and can provide both absolute and relative frequency scores for each service. Furthermore, daily absolute and relative usage for each service can be computed. To find out if learners have an account for an identified service, visits URLs of the service can be examined for keywords related to login/logout actions (e.g., 'login', 'logout', 'signin', 'signout', 'account'). However, this analysis only works as an indicator since these URLs might have been visited not only due to the user having an account for the respective service, but also due to a redirection to protect content from users that are not logged in. Further details on the implemented analysis can be found in [6].

Quality of Analysis

Since learners' visited the service's website, it can be ruled out that they do not know the respective service. However, this data is rather unreliable, as the automated extraction is limited to the learners' browser history. Based only on the entries in the browser history, it is difficult to tell whether learners are using a service actively or just visiting the website, or even if the website was visited by the current learner if access to the device is shared with other learners. Depending on how often learners visit a specific website or which particular resources they access, the collected data might allow for more analysis and further distinctions, but in the end, website visits

do not prove the use of a specific service. When multiple visits are detected, different paths and query parameters can reveal the potential use of the service. However, as services might be implemented as single-page applications, the browser history might only contain one entry per visit, even though multiple actions with the service have occurred. Furthermore, if different learners share one device (e.g., laptop or desktop computer) and access websites through the same browser, the available browser history does not reflect a single learner's interactions with different services. It potentially contains website visits of services the learner, who wants to play the personalized game, might not know. This creates a potential problem of reliability of the collected data and should be considered when analyzing the browser history.

Sensitivity of browser history

The browser extension for automated data collection presents a rather intrusive method of data collection in which learners grant total access to their private browser history. Therefore, learners' consent needs to be collected first before accessing any browser history entries. In addition to consent, handling information retrieved from learners' browser history is as important and needs to follow data protection regulations as the browser history data can be considered sensitive, personal data. As both games described in Chapter 6 are implemented using the MTLG framework, they are executed locally after their initial download from a web server. This allows for browser history to be used to personalize the games. By only executing the games locally, all data collected by the browser extension can be further processed locally as well, and thus, no sensitive data has to be sent to any external server. This way, the complete browser history and any analysis results containing partial browser history data or any personal information (e.g., which bank services a learner uses) do not have to leave the learners' device.

Problem with user studies

However, the approach is not yet suitable for large-scale user studies, as a potential study setup involves event logging within the game. While the automated data collector and further pipeline steps do not have to provide logging, services, and URLs used in gameplay will be logged for further analysis. Here, sensitive, personal information might be sent to a log server, which makes the option of using the automated data collector and complete event logging in the game controversial. For the evaluation in the scope of this work, the automated data collector will not be used. Before using the browser extension in the personalization pipeline, it should be evaluated under ethical considerations and regarding its acceptability among learners. The following questions of how automated data collection for learner modeling and personalization be designed transparently and in a way that is trustworthy to increase acceptability among learners, or how automated data collection based on learners' browser history can be used in a large-scale user study while supporting extensive logging functionality. Further, the question arises which information of the learners' browser history can be logged in a sensitive and privacy-preserving way. While this dissertation will not further elaborate on these questions, they remain for future work in the field of personalized anti-phishing learning games.

Learner modeling

The output of either data-collection approach provides the input to the learner modeling process. This step aims to generate a learner model, which is an abstract representation of the learner's characteristics, i.e., their online activity and familiarity with services. This model contains information about the services and websites a learner visits and is used as an initial knowledge base about the learner when creating the personalized learning game content.

Learner model

In the current implementation, the learner model is specified using a function object[36] containing a set of properties for relevant information about a learner and a set of functions to access different learner characteristics. This object stores the familiarity status of 'used', 'known' and 'unknown' services. The structure is highly flexible and extensible towards more information (e.g., bookmarked websites provided by the automated data collector) While the learner model is only based on previously collected information about the learner, future extensions could integrate a back-propagation loop in which gameplay data is analyzed, and useful gameplay information is stored in the learner model, e.g., performance scores or error rates. This allows for personalization through adaptivity since gameplay data can be combined with previously collected data to generate new levels adaptively.

As the learner model results from data collection, it presents an input to the following steps of the personalization pipeline. The learner model is mainly used in the content delivery module (in contrast to the initial concept, see 11). Since the level generator in the content delivery module accesses different content generators of the content generation module, the learner model is forwarded and provides input to different content generators.

7.2.2 Module for Content Generation

In order to provide personalized learning game content, game-specific content needs to be generated based on a learner model. Depending on the underlying game, the type of content can vary. As such, the module for content generation consists of various generator components. The learner model generated in the data collection module is provided as input and supports content generation by providing relevant learners' characteristics, i.e., familiarity with services. In the scope of this dissertation, the main generator component to provide personalized learning game content for the games presented in Chapter 6 is a URL generator. Other generators could be provided for different games, e.g., an email generator to personalize games on the topic of email phishing.

URL Generator

The objective of the *URL generator* is to provide large sets of different URLs, which can be integrated into the game prototypes of this dissertation. The component requires a list of services or URLs as input and is able to create custom sets of different benign and malicious URLs. Generated URLs follow a valid URL structure and contain additional URL parts, e.g., subdomain or path. They, however, do not have to be actual URLs that respective companies use to offer online services. In other words, the URL generator does not check for the validity of URLs by checking if they exist and responding to any kind of requests.

Rule-based URL generation

The input to the URL generator is a service name and its base URL, e.g., 'PayPal' and 'https://www.paypal.com'. Based on the configuration, the generator applies rule-based modifications to the URL to create a number of different URLs as well as details about their creation process, i.e., a protocol of applied rules. Several rules can be applied in sequence to create more complex and realistic URLs. The rules are based on patterns

[36] In JavaScript, functions are objects or 'action objects', which can be called but also be treated as objects (e.g., adding or removing properties). For more information, see `https://javascript.info/function-object` (accessed on 05.01.2022).

that were extracted from real-world examples of benign and phishing URLs: Benign rules are based on login page URLs of popular websites retrieved from Alexa[37], a web traffic analysis company that provides web traffic data and information about website rankings, e.g., the most popular websites worldwide. Malicious rules, however, are based on a study of related work (cf. [Rob+19; Agt+15; Kin+17; Oes+18]), and verified using PhishTank[38], a free, community-driven database on phishing data. As such, identified rules from related work were manually checked by analyzing reported phishing URLs in the PhishTank database. If a rule was found in reported URLs, it was considered still valid (not yet outdated), and considered for URL generation. Furthermore, reported URLs in the database were examined for additional rules, i.e., to find rules that were used in the wild but not yet reported in related work.

In its current implementation, the URL generator supports rules to generate different URL parts and, when combined, complete benign and malicious URLs. The generated URLs match one of the categories covered in the games (see Chapter 6.1.1). This means, only one part of the URL is actively manipulated by the rules and the resulting URL does only belong to one of the categories. For an overview of all implemented rules, see Table 36 in Appendix B.3.2. In addition, each generated URL is tagged with the respective familiarity level as stored in the learner profile.

On-demand URL generation

When generating URLs, the learner model is used as an input to create different pools of benign and malicious URLs for different levels of familiarity. Each URL is tagged with the respective level of familiarity according to the learner model. Depending on the progress in the game, different URL categories are required, e.g., for classification in the analysis game (see Chapter 6). As such, the URL generator is configured to generate URLs for all requested URL categories. In between levels, new URLs can be generated on demand to meet the requirements for the next level, i.e., benign and malicious URLs of a new category.

Further content generation

In the scope of this dissertation, only a URL generator was required to generate personalized URLs for the games described in 6. However, further content generators could be considered depending on the content a game requires. Since there are various types of phishing as introduced in Chapter 2.1.3, different generators can be implemented, e.g., for emails, websites, SMS or text messages.

Email generation and templating

For a personalized anti-phishing learning game on the topic of email phishing, personalized emails would be needed (e.g., if the game requires learners to decide whether a given email is legitimate or phishing). The generation of email can be done in two possible ways. On the one hand, methods for natural language generation can be used to generate fictional emails for different contexts, e.g., bank inquiries or account notifications. On the other hand, email templates could be used to rely on non-fictional email elements. As described in the concept for the personalization pipeline in Chapter 7.1, email templates for different categories (e.g., retail/shopping, financial services) could be used and filled with information from the learner model (e.g., name and base URL of used financial services or known online shops). URLs embedded in the email body could be generated using the URL generator described above. For the current version of the personalization pipeline, no email generator has been implemented yet.

[37] https://www.alexa.com/topsites, accessed on 09.12.2020
[38] https://www.phishtank.com/, last accessed on 09.12.2020

Generation of more complex content types

If game content is not only generated for different types of phishing, but for more complex game scenarios, content generation can be done using multiple steps. While the generation of phishing websites could be based on original benign websites, various opportunities exist to generate new content. As such, the generation of HTML, CSS, and JavaScript could be considered to alter the original website in aspects of structure, design, or functionality. Similar to email generation, the URL generator could be utilized to create benign and malicious URLs in the website code. For the scope of this work, no generator for phishing websites has been implemented. However, related work on phishing kits presents a good starting point for the design of a website generator (e.g., [CKV08]).

7.2.3 Module for Content Delivery

In order to embed personalized content into the games and provide a personalized version of the game, the module for content delivery provides two distinct components: The level generator and the level controller. While the pipeline was initially described as a three-step process, the content delivery module takes on the role of a controller, collecting necessary information with the help of several new components. The module provides interfaces to both the module for data collection and the module for content generation. It also provides an interface to the game that will be personalized.

Level Generator

As the output of the data collection step in the personalization pipeline, the learner model is provided as input to the *level generator*. The objective of the level generator is to create different levels depending on the current game state. The level generator uses the learner model as input to different content generators, and as a result, it receives generated, personalized content as explained in Chapter 7.2.2. Next, the generated content is embedded into an explicit level definition, i.e., a description format that provides suitable parameters and required level information for the game. Level definitions are tailored towards the specific game and include all information necessary to create a playable level, including specific task descriptions, conditions required to complete the level, and the URLs that are to be used. For the analysis game (described in Chapter 6), the level definition includes the list of URL categories (i.e., buckets) that are available in the level and for which personalized URLs were generated. The first version of the level generator was developed as part of a student thesis project [8]. For the scope of this work, it was reimplemented to fit the particular games and provide suitable level definitions with generated, personalized URLs.

Level Controller

The interface between the content delivery module and the game is implemented in the *level controller*. The level controller triggers the creation of new levels upon request by the game and returns appropriate level definitions to the game. This cyclic relation allows for on-demand level generation during the game. Finally, the game handles the translation of level definitions into actual playable levels by creating the required views and configuring the game logic.

8 Evaluation

This chapter covers the evaluation of both the learning game prototypes for anti-phishing education (described in Chapter 6) and a personalized version of one game prototype using the personalization pipeline (described in Chapter 7). The evaluation consists of two user studies following a pre-test/post-test between-group design as well as a longitudinal post-test to evaluate the long-term effects and self-reported behavior changes. As such, the first user study (see Chapter 8.1) presents a comparative evaluation of the analysis game, the creation game, and the decision game (i.e., the baseline implementation similar to related work). The second user study is an extension of the first evaluation, in which a personalized version of the analysis game is compared to its non-personalized version, and a gameplay analysis is performed to take a closer look at the learners' actions in the personalized game (see Chapter 8.2). To evaluate the long-term effects of the different game prototypes and personalization and self-reported behavior changes, a longitudinal study is presented in Chapter 8.3.

Large parts of this chapter have been previously published in [B], [I] and [J], and are based on the collaborative work with Vincent Drury, in which both participated equally. The research data collected and used for evaluation was published in [D]. However, this chapter contains a more comprehensive description of both user studies and the evaluation of long-term effects. It allows connecting them to each other and the overall scope of this dissertation.

8.1 Comparative Study of Different Learning Game Prototypes

In this section, a comparative user study using non-personalized implementations of the game prototypes described in Chapter 6 is presented. First, research objectives and questions are described in Chapter 8.1.1. Next, the user study is outlined with its design, the recruited participant sample as well as the used apparatus, materials and procedure. Also, a brief description of the game prototypes and their configuration (see Chapter 8.1.2). While Chapter 8.1.3 presents the results of the user study, a discussion follows in Chapter 8.1.4. The content of this section was previously published in [B], while the respective research was made available in [D]

8.1.1 Research Objectives and Questions

Differences between games

Based on the identified research gaps in the scope of this dissertation, the following evaluation allows for a comparison of the two new learning game prototypes with each other but also with a baseline implementation similar to existing work. While existing learning games most often implement a binary decision mechanic in which learners have to decide whether a given URL is malicious or benign, the new game prototypes use different game mechanics that can provide more insights into the learners' decision processes and possible misconceptions (see Chapter 5). As described in Chapter 6,

the analysis game implements an extended decision scheme in which learners have to analyze a given URL and sort it into different URL categories representing different manipulation techniques in Phishing URLs (e.g., 'Legitimate', 'Registrable Domain', 'Path'). However, the creation game takes a more constructive approach in which learners have to apply manipulation techniques to create their own malicious URLs. The comparison of the novel game prototypes can provide meaningful findings and extend the state of the art of anti-phishing learning games by showing the effectiveness of different, more complex game mechanics. Furthermore, the evaluation of the games allows further understanding the learners' performance in classifying different categories of phishing URLs and may detect possible misconceptions in the learners' understanding of the URL structure and different manipulation techniques.

Familiarity with services

In addition to comparing the game prototypes and understanding the decision processes of learners when classifying URLs, no personalization has been explored in anti-phishing learning games. In particular, familiarity with services and URLs displayed in anti-phishing learning games has not been considered so far, and existing studies did not consider whether the services used in the studies were actually known to participants or not. Therefore, this study focuses on participants' performance scores and confidence levels and considers the differences between services the participants use, know, or do not know since they can provide meaningful insights for the design and implementation of personalized learning games.

Research objectives and questions

For the evaluation of the three games and the particular aspects regarding the familiarity of services and possible differences between selected URL categories, a user study with three groups of participants was designed, where each group plays one game. This evaluation aims to evaluate the three games in a pre-/post-test study setup focused on URL classification knowledge. Therefore, the learning outcomes of all games are evaluated by comparing the participants' performances and confidences after playing either game. Furthermore, the effect of familiarity of services on classification performance is analyzed, and the initial state, improvements, and in-game behavior for classifying different URL categories are compared. To guide this evaluation, the following research questions (RQs) were formulated:

- **RQ-1**: Does playing 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? In particular, 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**: How is the participants' confidence in classifying URLs influenced by the games?
- **RO-5**: How is the participants' perception of phishing influenced by the games?

Beyond these RQs, another question regarding the potential differences in the participants' performance of classifying different URL categories is considered in the previously published paper [B]. The results and their discussion are, thus, not included in this work as they are partially out of scope.

8.1.2 Study Setup

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 anti-phishing learning games described in Chapter 6 serve as independent variables, and participants were assigned randomly to play one of the games. Additional independent variables were the chosen URLs and services in pre- and post-test and in the game. The performance and confidence in pre- and post-test served as dependent variables and were derived from the before mentioned research questions in Chapter 8.1.1.

Game Prototypes

In the scope of this study, the three different game prototypes described in Chapter 6 were used, i.e., 'All sorts of Phish' called *analysis game*, 'A phisher's bag of tricks' called *creation game* and the *decision game*, a baseline implementation of the analysis game with a binary decision scheme similar to related work.

URL Classification Test

To evaluate participants' performance and confidence in classifying URLs, a URL classification test was designed similar to tests used in related user studies. Therefore, a total of eleven categories of URLs are considered, consisting of one benign and ten malicious categories (see Table 12). This categorization enables a more detailed analysis of the participants' classification performance, as differences between the categories might indicate classes of URLs that are inherently more complicated to detect for users in our study.

Table 12: 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.

URL categories

The 'Benign' category includes all URLs that are considered benign in the context of the study, i.e., all URLs with an existing and legitimate *registrable domain*[39]. For the

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

categories of phishing URLs, the underlying assumption is that phishers target a particular original domain (e.g., 'ebay.com') and try to create phishing URLs that look as if they belonged to this original domain (e.g., 'ebay-service.com'). It can be differentiated whether a deceptive keyword appears in a *subdomain*, the *registrable domain*, the *path* (including queries and fragments) or not at all. Beyond these parts, manipulations of other URL parts are possible (e.g., authentication information or a port specification in the hostname) but were not included in the scope of this work, as they are less common, and in particular, did not appear on any benign login page that was encountered during the study design.

It should be noted that the URL categories described in Table 12 are presented and taught in a simplified way in the games to avoid confusion and reduce the amount of time that is required (see Chapter 6.1.1). 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. Further explanations of the different subcategories can be found in the previously published paper [B]. However, as the games rely on the grouped categories, these details are partially beyond the scope of this dissertation.

URL generation

The games, as well as the URL classification test used in the scope of this study, require example URLs, which were selected from a pool of benign and phishing URLs that was created as follows. The pool was constructed to create a representative set of phishing and benign URLs. As such, a set of relevant domain names was constructed by selecting services of various 'types' (e.g., shopping). To this end, the 50 most popular websites in Germany were selected (according to Alexa[40]). Next, 20 websites were removed in a manual review as they are either adult websites or websites whose landing page was not displayed in German 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 [APW21] and the 10 most commonly phished targets in Phishtank[41] (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[42] list which fit the type of service. In all, this results in 38 service names and their corresponding registrable domains, which are expected to be well known in Germany. The services are further extended by a URL that points directly to a login form, as determined by manually visiting each service's website.

The services are then used to generate the URLs that appear in the three games as well as the URL classification test as part of the pre- and post-test. This automatic generation is based on simple rule-based modifications of the input URL and results in 'Benign' URLs recognizable by their legitimate registrable domain but might not exist in the real world. The analysis and decision games randomly select examples from the remaining URLs and present them for learners to classify. Since the creation game only requires benign reference domain names, the registrable domains of the services were used here.

Test construction

To assess differences between URL categories, a URL classification test for use in pre- and post-test is created. The test consists of a binary classification task, with URLs selected based on the categories described above. For each URL, participants had to

[40] https://www.alexa.com/topsites/countries, accessed on 16.02.2021
[41] https://www.phishtank.com/, accessed on 16.02.2021
[42] https://tranco-list.eu/, accessed on 16.02.2021

decide whether it was benign or phishing and rate their confidence in the decision on a 6-point Likert scale (from 1 = 'very uncertain' to 6 = 'very certain'). 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 41 in the Appendix). The pre-test consists of 13 malicious URLs, which were selected by choosing example URLs from all categories.The seven 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 were 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. The URL classification test only includes URLs and 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 (e.g., [AAC15]).

In conclusion, the URL classification test measures the performance and confidence in classifying a set of URLs. The test contains a set of 20 different URLs and was utilized in both pre- and post-test to answer **RQ-1** to **RQ-5**. Ten additional URLs were added for the post-test to test for a possible learning bias. While an example URL is depicted in Figure 13, the complete list of all URLs used in both, pre- and post-test, is provided in Table 41 in the Appendix B.4.

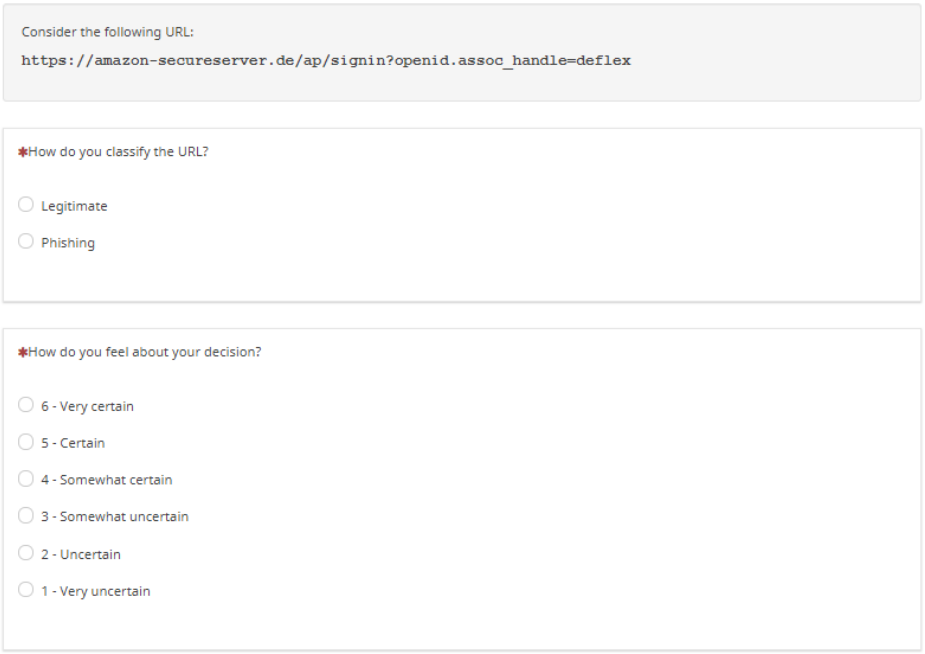

Figure 13: Screenshot of a URL in the URL classification test showing a malicious URL for participants to classify and rate their confidence in the decision.

Additional Materials and Apparatus

Remote lab study setup

The study was conducted in a remote, online lab study setup in which participants used their own devices to partake in the study. Due to the remote setup, participants

were required to use the video conferencing software Zoom[43] and either the web browser Google Chrome[44] or Mozilla Firefox[45]. In order to avoid possible technical issues with the used software, participants were asked to check for possible updates for the software before participating in the study.

Survey components

For the pre- and post-test phase, an online survey was provided using a self-hosted LimeSurvey application[46]. The survey consisted of the following questionnaires and tests:

- **Perception of Phishing**: This questionnaire is an adapted subset of the questionnaire used in [Ara12]. While the original questionnaire contained more items and was used in a different setting (i.e., not a user study using an anti-phishing learning game in a pre-test/post-test design), the evaluation revealed that the items for perceived threat were most influential besides perceived susceptibility and severity [Ara12]. Therefore the questionnaire was reduced, and the used version only included the three constructs of perceived threat, susceptibility, and severity. Some items were also reformulated to fit better in the context of a user study on anti-phishing learning games. The adapted version measures the participants' perception of phishing as a risk using a 5-point Likert scale (from 1 = 'strongly disagree' to 5 = 'strongly agree') and was used in both, pre- and post-test (see Table 13).

Table 13: **Perception of Phishing** questionnaire with three constructs: perceived susceptibility (PSU), perceived severity (PSE) and perceived threat (PTH). A german translation is provided in Table 37 in the Appendix B.4

ID	Questionnaire item
PSU1	It is extremely likely that I will be a victim of phishing in the future.
PSU2	My chances of becoming a victim of phising are high.
PSU3	I do not think that a phishing attack on me will be successful.
PSE1	A phishing attack would steal my personal information from my device without my knowledge.
PSE2	A phishing attack presents a risk to my privacy.
PSE3	I feel phishing attack would not steal my personal information from my device without my knowledge
PSE4	I don't think a phishing attack would threaten my privacy.
PTH1	Phishing attacks pose a threat to me.
PTH2	A phishing attack is a threat to my personal data.
PTH3	It is risky to use my device if I am the target of a phishing attack.
PTH4	I think that a phishing attack will not harm me.

- **URL classification test**: As introduced in Chapter 8.1.2, this test measures the performance and confidence in classifying a set of URLs. It is used in pre- and post-test.

[43] https://zoom.us/, accessed on 12.11.2021
[44] https://www.google.com/chrome/, accessed on 12.11.2021
[45] https://www.mozilla.org/firefox/, accessed on 12.11.2021
[46] https://survey.elearn.rwth-aachen.de, accessed on 11.12.2021

- **Recognition of Services**: This questionnaire contains a list of services that were targeted in the URLs of the URL classification test (usually embedded in the registrable domain of a URL, e.g. 'eBay' in 'https://www.ebay.de') and participants were asked to rate their familiarity with these services. Participants had to select whether they (a) use the service, (b) do not use but know the service, or (c) do not know the service (in response to **RQ-3**). The questionnaire covers all services, which are targets in any pre- and post-test URLs (see Table 42 in the Appendix for the complete list of services). It was as part of the post-test.
- **Demographics**: This questionnaire contains common questions regarding gender and age but also questions regarding the participants' educational background, experience with CS education, and self-reporting of prior knowledge in CS, IT security, and phishing. It was used in the post-test and was included to report potential biases among the participants. While in the first part of the study, self-reported prior knowledge was rated using a 6-point Likert scale from 'Very bad' (1) to 'Very good' (6), the scale was changed to an unbalanced scale from 'None' (1) to 'Very much' (6). Due to this change, the results are no longer comparable between groups, but they were still used to determine the prior knowledge of different participants. A complete overview of the demographics questionnaire is provided in Table 38 in the Appendix.

Survey structure

As such, the pre-test part of the survey first contains the **Perception of Phishing** questionnaire and afterward the **URL classification test** with 20 pre-test URLs. Similar to the pre-test, the post-test was structured as follows: (1) **Perception of Phishing**, (2) **URL classification test** with 20 pre-test URLs and 10 additional URLs, (3) **Familiarity with Services** and (4) **Demographics**. In addition to the above pre- and post-test, the three game prototypes described in Chapter 6 were used as central experimental interventions. The online survey and the games were provided online and accessed via the participants' web browser.

Procedure

The user study was structured using multiple sessions divided into seven different phases: (1) participant arrival, (2) briefing, (3) pre-test, (4) playing, (5) post-test, (6) debriefing, and (7) session closing. Each session was hosted by two instructors and included at most 10 participants. All participants were asked to log in at least five minutes prior to the session's starting time so that technical issues could be resolved before starting the session. Since the instructors were present at all times, participants were asked to activate their microphones to request help from one of the instructors.

Introductory information

In the briefing phase of a session, the procedure of the user study as well as the requirements for participation were explained, and participants had the chance to ask any open questions before participating in the user study. To give more contextual information and establish a shared understanding, a definition of phishing, including an example, was presented (see instructional information in Listing 1 in the Appendix).

Before continuing with the pre-test phase, participants were instructed to continue through the subsequent three phases individually and at their own pace. The participants were asked to inform the instructors when they completed the post-test phase. No other communication between participants and instructors was needed until the debriefing phase. After completing the pre-test phase of the study, participants received access to the game in a second browser tab. The final screen of the games instructed participants to return to the survey tab and continue with the post-test phase. When

all participants finished the post-test phase, the instructors explained the purpose of the study and answered any remaining questions by the participants in a debriefing before closing the session. Participants who completed the study early were asked to remain calm and wait until the study's debriefing phase.

Randomization

The decision of which game was to be played by which participant was made uniformly at random by the survey software when each participant started the pre-test phase. Participants did not know that different games were tested in the study or which group they were randomly assigned.

Ethical review and data protection

It should be noted that RWTH Aachen University does not have an ethics committee that could have approved this study. Instead, the study was designed similarly to existing studies that ethical committees approved. It complies with strict data protection policies, as discussed with the data protection officer of the university.

Participants

The study was conducted in two parts. While the first part was conducted in November 2020 with 88 participants who played the analysis and creation game, the second part was done in May 2021 and included 45 participants who played the decision game. Only the game was switched to the decision game between both parts, and the whole study setup remained the same. Potential differences within the population can be neglected, as the topic was potentially new for every participant, and no immediate increase of knowledge among the general public was suspected.

Recruiting

Recruiting was done online by posting information about the study and important instructions on how to participate in different social network groups of universities and distributing it via university mailing lists. Recruitment focused on people interested in learning about IT security, regular online activities, and little to no prior knowledge in IT security and CS. However, none of these aspects was checked explicitly prior to participating, and everyone interested was allowed to participate in the study. Since the study required active participation and availability for 60-70 minutes, a financial incentive of 15 € was offered to each participant.

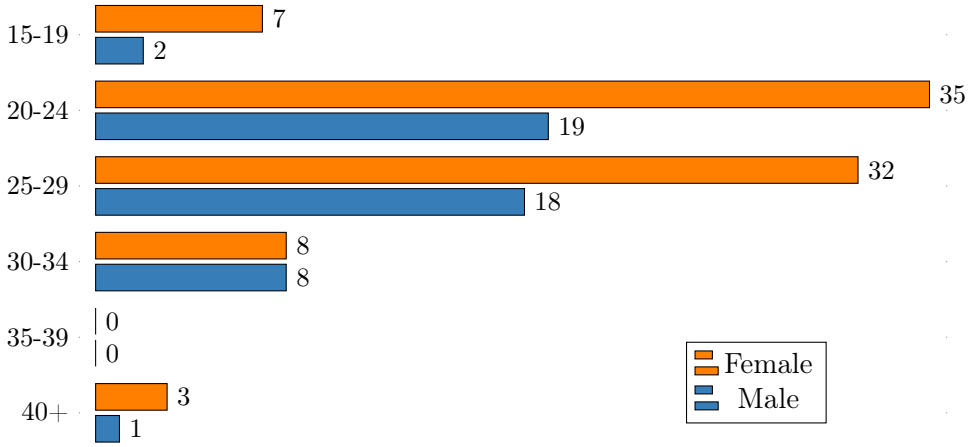

Figure 14: Distribution of Age and Gender ($N = 133$)

Demographics

Among the total sample of participants, 85 identified as female (63.91%) and 48 as male (36.09%). At the time of the study, most participants were between 20-29 years old (78.20%). Figure 14 shows the distribution of age and gender. Due to the recruiting methods, 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 have mainly completed their studies with a Master's degree (12.03%) or have completed vocational training (3.01%).

Exclusion of participants

Besides the total of 133 participants, an additional five participants were excluded from the analysis 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.

8.1.3 Results

Based on the research questions described in Chapter 8.1.1, a series of analyses and tests was conducted and will be reported in the following. For each test, three different groups were considered depending on which game the participants played: the *analysis game group*, the *creation game group*, and the *decision game group*. Collected research data can be found in [D].

Measurements and notations

In general, the participants' performance scores in the URL classification are computed as relative scores, i.e., the number of correctly classified URLs divided by the number of all URLs. Similarly, the confidence levels are equal to the average confidence ratings (ranging from 1 to 6) in either pre- or post-test. Performance scores and confidence levels are measured using an interval scale. For testing, a significance level $\alpha = .05$ was used. All statistical tests were computed using JASP[47]. Furthermore, the notations listed in Table 14 were used for variables computed during statistical testing. Note, that 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*). Also, indices **A** (for analysis game), **C** (for creation game) and **D** (for decision game) are used to indicate the respective game group.

Table 14: Notations for statistical tests

Variable	Meaning
M	Mean
SD	Standard deviation
α	Significance level
p	Significance value of statistical test
t	Test statistics for Student's t-test
W	Test statistics for Wilcoxon signed-rank test
F	Test statistics for ANOVA
ϵ	Greenhouse-Geisser estimate
d	Effect size by Cohen
r	Rank-biserial correlation coefficient
η_p^2	Partial η^2 estimates of effect size for ANOVA
ω	McDonald's reliability coefficient

[47] https://jasp-stats.org/ online, accessed 2021-08-26

Evaluation of potential learning bias

Before addressing individual research questions, potential learning bias was checked by comparing the performance means in the post-test (see Table 15). 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 than the means of overall post-test (M_{post}). As such, the effect of learning bias is negligible. While only the URLs that were part of the pre- and post-test are considered in answering **RQ-1**, all URLs used in the post-test are considered for further research questions.

Table 15: Means and standard deviations for performance scores (relative scores) in pre- and post-test including means on partial URL sets

Game	N	M_{pre} (SD)	$M_{\text{post-pre}}$ (SD)	M_{post} (SD)	$M_{\text{post-new}}$ (SD)
Analysis	40	.695 (.098)	.828 (.115)	.840 (.095)	.853 (.140)
Creation	48	.702 (.122)	.755 (.122)	.782 (.129)	.838 (.163)
Decision	45	.701 (.097)	.818 (.091)	.831 (.097)	.858 (.141)

Differences Between Pre- and Post-Test

The first research question (**RQ-1**, as described in Chapter 8.1.1) focuses only on the general effectiveness of the games, i.e., whether playing the games has a positive influence on the participants' performance in classifying URLs. As such, the following hypothesis was derived: The participants' performance in classifying URLs increased after playing either one of the games.

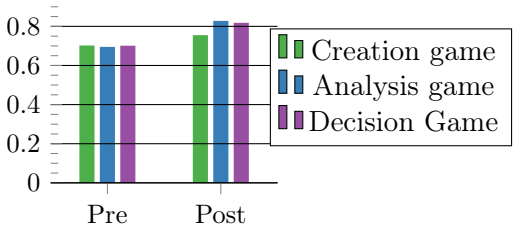

Figure 15: Differences between mean pre- and post-test relative performance scores for both games on pre-test URLs ($\mathbf{H1_{a/b}}$)

As shown in Figure 15, the mean performance score improved in the post-test for all three games with its highest value for the analysis game group. For further comparisons, Table 15 contains all relevant mean performance scores of the different game groups.

For statistical testing 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. If a deviation from normality was detected (Shapiro-Wilk test, cut-off value $\alpha < 0.05$), the non-parametric Wilcoxon signed-ranked test was performed instead of a t-test. The results of statistical testing (see Table 16) indicate that the participants' performances increased significantly for all three games. There are, however, differences in effect sizes, which are large for the analysis game group and the decision game group but not for the creation game group.

Differences Between the Three Games

Regarding **RQ-2**, the following hypothesis was formulated: The participants' performance in classifying URLs in the post-test differs between the three games.

Table 16: Results of t-tests comparing performance scores in pre- and post-test for all three games

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

*Deviation from normality detected.

First, comparing the mean values (see Table 15) suggests that participants of the creation game group performed worse in the post-test than participants of the analysis game group and the decision game group, who, however, performed similarly well. Next, statistical testing was used to evaluate the differences further, and as such, performance scores in the post-test were compared (i.e., performance on all URLs of URL categories that were part of all three games, including post-only URLs). Post-only URLs were included, as a higher number of URLs gives a more precise measurement of the participants' performance. An ANCOVA (analysis of covariance) was performed with the different games as the between-subject factor, performance scores in the post-test as the dependent variable, and performance in the pre-test as the covariate. It should be noted that Levene's test was not significant ($F(2, 130) = 1.207, p = 0.302$), meaning that all three game groups have equal variances. The ANCOVA does not return significant results for the three game groups as between-subject factor ($F(2, 129) = 0.505, p = 0.605, \eta_p^2 = .008$), only for the pre-test score as covariate ($F(1, 129) = 45.333, p < 0.001, \eta_p^2 = .260$). Thus, the hypothesis that there are significant differences in post-test performances between the games is rejected. Of particular note is that the more complex sorting mechanism included in the analysis game did not result in significant differences to the decision game.

Differences Between Used, Known and Unknown Services

For **RQ-3**, the focus lies on the possible differences between services that the participants use or know compared to services they do not know (see Table 42 in the Appendix for an overview on covered services and the absolute and relative familiarity levels). As such, the following hypothesis is being tested: The participants' performance in classifying URLs of services they use or know is better than for services they do not know.

Table 17: Performance scores (and standard deviations) per service familiarity

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

First, descriptive statistics show higher performance scores for used and known services compared to unknown services (see Figure 16 and Table 17). To test for significant differences in performance scores, two repeated-measure ANOVAs (pre- and post-test) were performed using the relative performances for URLs of different levels of knowledge as the repeated-measures factor. The games were used as a between-subject factor in

the post-test but not in the pre-test since no differences between the groups are assumed before playing the games. For both ANOVAs, Mauchly's sphericity test was significant, thus violating the assumption of sphericity ($p <= 0.05$). As such, the degrees of freedom were corrected using Greenhouse-Geisser estimates ($\epsilon_{pre} = 0.951, \epsilon_{post} = 0.750$).

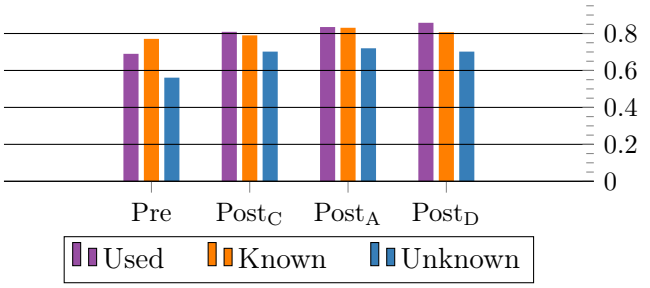

Figure 16: Differences between relative performance scores for used, known, and unknown services for combined pre-test, as well as post-test of all games

In the ANOVA performed on pre-test performance scores ($N = 115$), familiarity with a service had a significant effect on the participants' performance ($F(1.902, 216.815) = 19.643$, $p < 0.001$, $\eta_p^2 = .147$). Furthermore, post-hoc tests using Holm's correction confirm significant differences between unknown and known or used services: unknown and known ($p < .001$, $d = -0.539$), unknown and used ($p < 0.001$, $d = -0.466$). No significant differences were observed for known and used services ($p = 0.432$, $d = 0.073$).

Similar results can be observed for the ANOVA using performance scores in the post-test ($N = 122$, $F(1.500, 178.552) = 16.227$, $p < 0.001$, $\eta_p^2 = .120$). Here again, post-hoc tests using Holm's correction only indicate significant differences for unknown URLs: unknown and known services ($p < 0.001$, $d = -0.385$) as well as for unknown and used services ($p < .001$, $d = -0.490$). The differences are not significant for known and used services ($p = 0.246$, $d = -0.105$). It should be noted that effect sizes are not adjusted for multiple measurements in the post-hoc tests.

Overall, familiarity with a service affects the participants' performance in pre- and post-tests, with significant differences for unknown URLs compared to known and used URLs. Differences between services that are known (but not used) and used services, on the other hand, did not result in significant effects.

Effects on the Participants' Confidence

Next, effects on the participants' confidence are evaluated in response to **RQ-4**. The evaluation focuses on possible differences between pre- and post-test and between the games.

First, the mean confidence levels for pre- and post-test in each game group were computed (see Table 18). While there are overall differences between mean confidence levels in the pre- and post-test, differences between means for different game groups are less evident. Only for the creation game group can it be observed that the M_{post} and $M_{\text{post-new}}$ are smaller compared to the means of the other games. This indicates more minor confidence improvements for the creation game group than for the analysis game group or decision game group. As such, the participants' confidence levels seem to mirror their performance scores, as they increase after playing the games, with participants of the creation game seeming to feel less confident after playing compared to the analysis and decision games.

Table 18: Means and standard deviations for confidence levels (range: 1-6) in pre- and post-test including means on partial URL sets

Game	N	M_{pre} (SD)	$M_{\text{post-pre}}$ (SD)	M_{post} (SD)	$M_{\text{post-new}}$ (SD)
Creation	48	4.118 (.720)	4.701 (.625)	4.751 (.642)	4.923 (.723)
Analysis	40	4.065 (.637)	5.034 (.468)	5.086 (.461)	5.065 (.764)
Decision	45	4.129 (.714)	5.004 (.542)	5.068 (.500)	5.113 (.580)

In the next step, statistical testing was performed to evaluate the potential significance of existing differences. Here, the results show significant improvements in the participants' confidence levels between pre- and post-test with a large effect size in all games (see Table 19). For the creation and decision games, parametric t-tests were used, but due to a deviation from normality (Shapiro-Wilk test, cut-off value $\alpha < .05$), a non-parametric Wilcoxon signed-rank test was used for the analysis game. Next, a one-way ANCOVA was conducted to determine statistically significant differences in the participants' confidence in the post-test between the three games while controlling for the participants' confidence in the pre-test. This was done using the participants' performance scores on URL categories that were part of all three games. The results indicate significant differences between the three 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 ($F(1, 129) = 79.372$, $p < .001$, $\eta_p^2 = .381$). Furthermore, post-hoc testing using Holm's correction showed significant differences between the creation game group and analysis game group ($p = .015$) as well as the creation game group and decision game group ($p = .015$) but not between the analysis game group and decision game group ($p = .892$).

Table 19: Results of t-tests comparing confidences in pre- and post-test for all three games

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

*Deviation from normality detected.

Effects on the Participants' Perception of Phishing

For **RQ-5**, the following hypothesis was derived: The participants' perception of phishing is different after playing either one of the games. To test this hypothesis, the perception of phishing questionnaire is used in the pre- and post-tests. Computing McDonald's ω, the reliability of the questionnaire is acceptable in the pre-test ($\omega = .730$) and good in the post-test ($\omega = .870$). Note, while initially the questionnaire was divided into three constructs (see Chapter 8.1.2), evaluation of reliability revealed that the joint construct of the participants' perception of phishing as a risk provides more reliable measurements. As such, the participants' answers to all questionnaire items are combined into a single score (using the mean value), which is then analyzed for changes between the pre- and post-test for all three games.

Comparing mean values, a decrease between pre- and post-test can be observed for all three groups. The results of statistical testing show significant differences in the

analysis game group and the creation game group, while the test for the decision game group was inconclusive (see Table 20). A deviation from normality was observed for the analysis game group and the decision game group (Shapiro-Wilk test, cut-off value $\alpha < .05$). Thus, the Wilcoxon-signed rank test was used instead of a t-test.

Table 20: Mean values and results of t-tests comparing relative scores for perception of phishing questionnaire in pre- and post-test for all three games

Game	M_{pre} (SD)	M_{post} (SD)	Test statistic	p-value	effect size
Analysis*	3.870 (.512)	3.577 (.679)	$W = 526.500$	$p = .002$	$r = .581$
Creation	3.720 (.574)	3.494 (.785)	$t(47) = 2.615$	$p = .012$	$d = .377$
Decision*	3.737 (.483)	3.549 (.714)	$W = 601.000$	$p = .124$	$r = .271$

*Deviation from normality detected.

In response to **RQ-5**, the derived hypothesis can be accepted for the analysis game and the creation game. However, the hypothesis needs to be rejected for the decision game group since no significant differences were observed between the pre- and post-test.

8.1.4 Discussion

While in the previous section, the results of the user study were outlined in response to the initially formulated research questions (described in Chapter 8.1.1), this section will present a discussion of the findings and possible limitations. Furthermore, future work directions and implications for a follow-up study are discussed.

Summary of results

The results of the user study include significant differences in the participants' performance after playing either one of the games (**RQ-1**), but no significant performance differences between the three games (**RQ-2**). Regarding the familiarity with services (**RQ-3**), participants were significantly better in classifying URLs of services they know or use (compared to unknown services). In addition to the effects on the participants' performance, results show significant improvements in the participants' confidence levels after playing either one of the games as well as a significant difference between the creation game group and the decision and analysis game groups (**RQ-4**), i.e., in the post-test, participants of the creation game were less confident then participants of the other games. Lastly, the results of the perception of phishing questionnaire show significant differences for the analysis game group and the creation game group, but not for the decision game group (**RQ-5**).

Study Setup

First, the chosen study setup, including its design and the selected participant samples, is discussed as decisions before or during the execution of the study can introduce certain limitations as well as influence the findings. Also, reviewing the study setup may lead to valuable insights for future work.

Generalizability

Taking a general look at the participants of the user study reveals a deviation from the general population. While recruiting was not explicitly limited to a specific age group or occupation, the advertising strategy for the study focused on online social groups for students and adult users of social networks (e.g., distribution via Facebook groups). As a result, the participant sample consists mainly of students and does not represent the general population, which might lead to problems in generalizing the study's findings. It

might be possible that younger people, such as students, have a different state of mind concerning online risks (cf. [Oli+17]) or have more experience in dealing with URLs than the general population. However, since most participants consist of students between 20-30 years old, results might be generalizable to this demographic. This claim could be substantiated by replicating the study and providing more supporting evidence. Additional user studies with more variety in the participant samples are recommended for further generalization to a larger demographic.

Effects on awareness

The presented user study was performed as a remote online study, where participants utilized their own, familiar devices to access the survey and game as well as log into the supervised video conference. Even though using their own devices in their own home are often seen design characteristics in field studies, the study design was a lab study and did not test how participants would respond to actual phishing attacks in a more realistic setting (e.g., through simulated phishing attacks as part of phishing campaign [VSB20]). Since the focus of the games is to impart the knowledge required to understand the basic structure of a URL and detect phishing URLs, the study shows how well this knowledge can be applied in an optimal setting where the participants are fully aware of the task. In particular, no claim is made that the games raise situational awareness or help avoid phishing attacks in a real-world setting since the games do not convey the knowledge and awareness of how and when phishers lure potential victims into disclosing personal information. This is beyond the current scope of the games.

Limitations of URL classification tests

In the pre- and post-test phases of the user study, participants were asked to classify a set of URLs (as part of the URL classification test). This did not include screenshots of potential websites. As it has previously been shown that users do not usually look at the URL bar, even in phishing classification tasks (see, e.g., [AAC15]), the URL classification test used in the study did only include URL strings and no additional information. As the focus of the games, and subsequently the complete study, is the knowledge about the URL structure and phishing URLs and how this knowledge can be utilized to detect phishing websites, the decision was made to only include URLs in the tests. Furthermore, imparted knowledge about URLs can be applied in several different contexts (e.g., to analyze URLs before clicking on them or understanding TLS certificates [DM19]). Therefore, the chosen URL classification task maps the requirements of the user study more precisely than a website classification task. However, it might be possible that effects for unknown services (as evaluated for **RQ-3**) are increased, as website screenshots would offer more context information in the decision process. To prevent this, the crafted URLs in the URL classification test always included a reference to the original service name (e.g., 'Otto' for 'https://b1ovam5.org/otto.de/'), and additional information of the website would therefore not have made a significant difference. The only exception would be random URLs, which do not include recognizable service names. Note, however, that random URLs were not included in the evaluation of **RQ-2** and **RQ-3** (see Table 41 in the Appendix for a complete list of all URLs used in the URL classification test). Finally, it should be noted that the tests also include more malicious URLs than benign URLs, which does not realistically reflect the real-world setting and might have led to bias in the results. This, however, does not seem to have an impact on the study, as the improved results for benign URLs (see Table 43 in the Appendix) in all three games indicate that participants were not simply choosing 'Phishing' more often, and did instead utilize their understanding of URLs to classify URLs more effectively. Furthermore, confidence improvements might indicate that participants also felt like they could apply their knowledge more effectively, thus deciding more confidently.

Long-term effects and alternative study designs

Furthermore, the study setup only assessed the performance and confidence levels over a short period of time (as the study follows a pre-/post-test design). It might be interesting to explore how the test results change over time and whether imparted knowledge was retained. As such, the study design was extended by providing the participants the chance to partake in a longitudinal study, in which a second test was distributed a distinct time period later to test for long-term effects. This extension of the study is presented in Chapter 8.3. As an alternative to evaluating long-term effects in a lab study setup, a longitudinal study, possibly supplemented by a simulated phishing attack, could provide further results. Here, aspects of awareness and behavioral change could be linked to actual phishing attack recognition when participants' are targeted with simulated phishing attacks. Since the current study setup was only designed in cooperation with the data protection officer of RWTH Aachen University, an ethical review is recommended to ensure good scientific practice and assess the impact and potential risks of the study.

COVID-19 pandemic

Also, it should be noted that the study was performed during the time of the COVID-19 pandemic, which might have had an impact on the participants' abilities and state of mind. However, the study did not test for the effects of the pandemic, and the assumption was made that the pandemic has not had a more significant influence on the results than other possible limiting but unknown factors.

Overall, the chosen study setup and participant sample have introduced different limitations to the results, which immediately affects generalizability as well as future studies. For the scope of this work, the study setup is not changed for the follow-up study presented in Chapter 8.2. Introducing different instruments, adapting the general setup, or changing recruiting approaches would interfere with the comparability of both studies. However, for future work beyond the scope of this dissertation, recommendations include more diverse participant sampling as well as the exploration of more realistic study setups (e.g., using simulated phishing attacks to test long-term effects).

Study Results

In this section, the results of the user study are discussed following the formulated research questions described in Chapter 8.1.1. For each aspect that was evaluated in the study, the results are interpreted, and limitations, as well as future work implications, are discussed.

Differences between games and problems of the creation game

As described in Chapter 8.1.3, results show significant increases in both performance scores and confidence levels from pre- to post-test in all three games (**RQ-1** and **RQ-4**). When comparing the three games (**RQ-2**), the participants' performance did not differ significantly, and only participants of the creation game group were significantly less confident in the post-test (**RQ-4**). To this end, it should be noted that some participants who played the creation game took more time and asked more questions during the study, as some of them had problems advancing through the games and thus, might feel less confident in their abilities. Another explanation for the difference in the participants' confidence is that the requirement to create URLs by themselves posed a higher difficulty and complexity, which may have confused participants who did not understand the learning content completely. Furthermore, it is noteworthy that the creation game differs from the analysis game (and the decision game) in the included URL categories. While in the creation game learners started with manipulation of the TLD, one of the first categories in the analysis game was the category 'IP address', in which URLs belong that use an IP address instead of a readable hostname. Furthermore,

the overall game design differs, as the analysis and decision game are set in a Roman theme and the creation game is more abstract (see Chapter 8.1.2 for a brief overview or Chapter 6 for details on the games). As such, the comparison is less significant than the comparison between the analysis game group and the decision game group. Since this presents a limitation to the study, for future work, the differences between the games should be addressed by aligning both games to improve the comparability of the games.

Advantages of the analysis game: insights into learners' decision processes

Beyond the comparisons with the creation game group, the comparison of the analysis game group and the decision game group did not yield significant differences in the participants' performance and confidence (**RQ-2** and **RQ-4**). This might be an effect of alignment, as it has previously been hypothesized that the alignment of game mechanics with the actual test can have an influence on the participants' performance 5.3, and the URL classification tests used in the pre- and post-tests of the study required a binary decision. Alternatively, it might indicate that more complex decisions (as implemented in the analysis game) do not actually make a difference. Still, it can be argued that the analysis game offers several advantages. For one, when performing an analysis of in-game log data for both games, the analysis game is able to offer more insights into learners' decision processes. This is because the learners' mistakes can offer a better understanding of possible misconceptions when URLs are sorted into buckets of different URL categories instead of making a binary decision. In the analysis game, learners have to decide which manipulation technique was applied and which URL part was manipulated in order to sort the URL into the correct bucket. Misconceptions can be detected if learners sort URLs of a particular category into wrong buckets repeatedly (e.g., classifying 'https://ebay.de.shop.eu' as 'RegDomain' can happen if learners have issues identifying manipulation of subdomains). As such, figures 17 and 18 shows the sorting outcomes of the analysis game and the decision game (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). Even though general trends are visible in the decision game as well, the choices and confusions of participants are more evident in the analysis game. While the data of the decision game group allows evaluating how much was classified incorrectly by the participants, the data of the analysis game group presents the alternative choices the participants made. The results show, for example, that participants of the analysis game group fail to classify URLs of the 'Path' category and sort them most often into the 'Random' category, which may indicate a possible problem in understanding the difference between these categories. This trend is not visible in the decision game, as the log data of the decision does not allow for this distinction and may indicate that participants focus mainly on the domain name for classification while mostly ignoring path information. Moreover, it is worth pointing out that the analysis game is better suited for understanding misunderstandings and problems in the learners' basic parsing capabilities of URLs. This can be seen, for example, with URLs in the subdomain category, which were often confused with RegDomain URLs. In the scope of this study, the results of comparing game log data are not yet fully exhausted and could be further explored in future work.

Familiarity with services

For **RQ-3**, differences between known and unknown services in the post-test were analyzed, and results show significant performance differences among all three games. Results show that the participants' performance on URLs of unknown services is significantly lower than performance on URLs of known or used services. Even though the performance increased for unknown services between pre- and post-test, they were not at the same level of known or used services (in the post-test). This indicates an overall problem in classifying URLs of unknown services as participants have no point of reference, do not know the correct decision, and therefore divert to guessing. Similarly, the

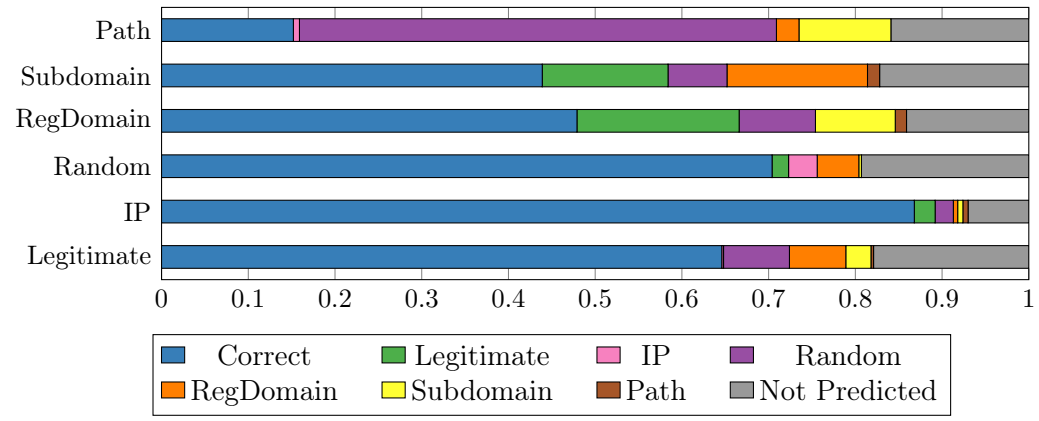

Figure 17: Relative sorting outcomes for each URL category in the analysis game

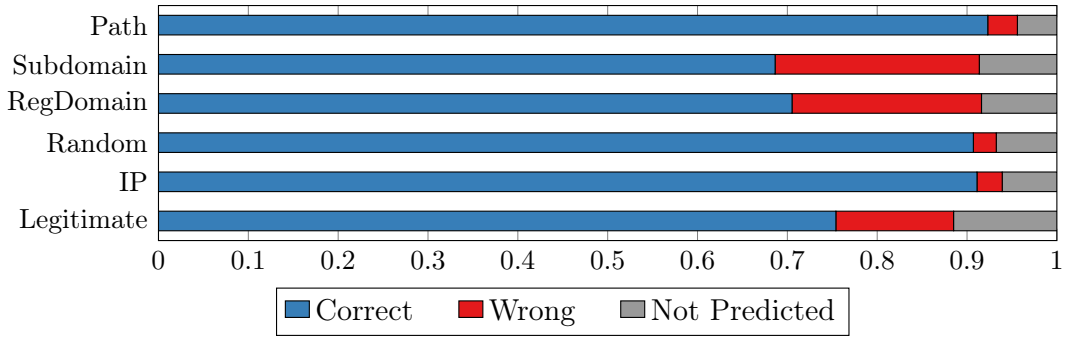

Figure 18: Relative sorting outcomes for each URL category in the decision game

results for participants' confidences mirrored the results for participants' performance scores. Participants are less confident in classifying URLs of unknown services compared to URLs of known or used services. These results open up 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. A limiting factor to these results is, however, the fact that there were generally fewer URLs with services that participants did not know than those they did know (3.26 unknown on average), with a few participants even knowing all the services, which might have introduced a bias in the comparison. In the future, it is recommended to include some lesser-known services in URL classification tests (and possibly the games) to assess potential differences in more detail and provide possibly more supporting evidence to strengthen the research in this field.

Perception of phishing as a risk

The results in response to **RQ-5** show that the participants' perception of phishing as a risk decreased between pre- and post-test, and the observed differences were significant in the analysis game group and the creation game group. At first, this may indicate that the two new game prototypes are more effective in raising awareness and lowering the fear of phishing as the perception decreases. However, the interpretation of these results should be done carefully since, on the one hand, the remote lab study setup does not provide a realistic setting and does not test for situational awareness. On the other hand, the questionnaire results might be influenced by a social desirability bias, as participants may feel the need to answer more in favor of the effectiveness of the games and their raise of awareness at the moment of the lab study. Also, the used version of the questionnaire was adapted and shortened compared to its original questionnaire (cf. [Ara12]). There is no data on this questionnaire in similar studies and no evaluation

of its overall quality as an instrument. An improved version of a perception of phishing questionnaire could possibly provide more reliable and valid results. Recommendations for future work, therefore, include the replication of the study with a larger participant sample but also a redesign of the questionnaire, possibly with domain experts and principles of psychological test theory.

Summary of discussion and future work opportunities

Overall, the results of the presented user study indicate the general effectiveness of the three game prototypes, with some differences in the effects of participants' performance and confidence in URL classification tests used in the pre-test/post-test design. While the participants' performance generally improved after playing either game, the results show differences between the three games, which may originate from differences in the participant sample or differences between the covered learning content and used game mechanics in the games. As such, future work calls for the reproduction of this study with larger participant samples, an alignment of the games, and possibly even the development of a combined version of the new game prototypes, i.e., a combined game that incorporates elements of the analysis game and the creation game. This way, advantages of both games could be utilized, and possible interaction effects on the participants' abilities could be evaluated, e.g., by comparing it in a similar study setup to the results of the presented user study. When reiterating over the design of the games, the use of an explicit game design framework in combination with BRT could be beneficial to the overall quality of the games. Another direction for future work lies in the exploration of service familiarity in URL classification tests but also within the actual games. Results of this user study show that participants performed significantly worse when classifying URLs for unknown services, which raises questions about the reproducibility of studies that do not consider the participants' familiarity with services used in their tests, but also about how incorporating service familiarity in the games could provide more insights into participants' abilities and problems in classifying URLs. This is directly related to the objectives of this dissertation, and therefore a follow-up study is presented in section 8.2. Finally, a test in a more realistic setting, with a more comprehensive set of educational resources and simulated phishing attacks, and an evaluation in a longitudinal study might be insightful additions in future work. In the scope of this dissertation, long-term effects were evaluated in a small longitudinal study which is presented in Chapter 8.3.

8.2 Comparative Study using Personalization and Gameplay Analysis

In this section, a follow-up user study using a personalized implementation of the analysis game (described in Chapter 6) is presented, and the results are compared to the sample of the initial user study. As such, research objectives and questions are described in Chapter 8.2.1. Regarding the overall user study design, it should be noted that only minor adaptations were made to be still able to compare participant samples (see Chapter 8.2.2). While the results of the follow-up user study are described in Sections 8.2.3 and 8.2.4, a discussion is presented in Chapter 8.2.5. The study and its results were previously published in [I], while the respective research data was provided in [D].

8.2.1 Research Objectives and Questions

Based on the identified differences between the different game prototypes compared in the initial user study (described in Chapter 8.1), a follow-up user study was conducted

to understand further how familiarity with services could be utilized in game-based anti-phishing education.

Service familiarity and its influence on participants' performance

As the results of the previous study show, the participants' familiarity with services does influence participants' performance. Results indicate a correlation between service familiarity and the participants' classification performance on URLs of the respective service. This can be seen by the significant differences regarding the participants' performance of classifying URLs of unknown and known or used services (described in Chapter 8.1.3). For this follow-up study, the focus lies on utilizing service familiarity within the games to adapt gameplay and further explore how learners handle URLs of services with different levels of familiarity. As such, a personalized version of the analysis game was created using the personalization pipeline developed in the scope of this dissertation (see Chapter 7.1). The personalization pipeline allows for dynamic URL generation based on a learner model containing relevant information of the learners' familiarity with different services. The generated URLs can then be embedded in the analysis game, and learners are required to classify them (similar to the first study since the game design did not change). In the following user study, the personalized game is compared to the non-personalized version of the analysis game, and the participant sample of the first user study is analyzed in comparison to the newly retrieved sample.

Since the creation game was outperformed by both the analysis game and the decision game (see Chapter 8.1.3), it was decided not to consider it in the follow-up study. The analysis of the creation game resulted in different issues, content-wise and with respect to the results. For one, the game was not entirely comparable to the analysis game and decision game, but also, the game was perceived as more difficult since participants asked more questions during the study when they had issues advancing in the game (discussed in Chapter 8.1.4). Yet, another argument for why the creation game was not considered for the follow-up user study is its limited potential for personalization. The creation game requires learners to combine given URL parts to create malicious URLs by applying different manipulation techniques, and each level of the creation game consists of two to three presets (i.e., tasks in which a URL has to be created). Thus, the number of tasks learners have to complete is rather limited, and personalization of the game's content does not allow for significant adaptations since only the task description and targeted URL with its underlying service could be personalized. Personalization would be minimal, and the potential effects on learners would be hard to attribute to these minor adaptations. Conclusively, the creation game was not considered for personalization and excluded from the follow-up user study.

Research objectives and questions

To understand the potential differences and advantages of personalized gameplay, the objectives of the follow-up study are two-fold. On the one hand, the objective is to understand the differences between the analysis game and the personalized version and how it affects the participants' performance and confidence. On the other hand, the personalized version allows for an even closer look at the learners' actions within the game, and as such, the objective is to analyze the gameplay and see how personalization influences the learners' in-game behavior. In conclusion, the following RQs were derived:

- **RQ-1**: Do the games, in particular the personalized version of the analysis game, have a positive influence on the participants' performance in classifying URLs?
- **RQ-2**: Are there differences in the participants' performance and confidence between the analysis game and the personalized version?
- **RQ-3**: How does personalization affect the participants' performance and confidence in classifying URLs?

- **RQ-4**: How does personalization (i.e., familiarity with services) affect in-game behavior?

It should be noted that while in the previously published paper [I], possible long-term effects were evaluated in a longitudinal test, only the pre- and post-test evaluation and the analysis of gameplay of the personalized game are described in the following. Results of longitudinal testing for the follow-up study and the initial user study are described in Chapter 8.3.

8.2.2 Study Setup

For the comparison of a personalized and a non-personalized version of the analysis game, a between-group design in a pre-/post-test setup was chosen, similar to the setup of the first study where three different game prototypes were compared (see Chapter 8.1). While the first participant group, which played the analysis game, took part in the study conducted in November 2020, the personalized version was tested with a participant sample in May 2021. Both versions of the game serve as the independent variables, while the participants' performance and confidence in the pre- and post-tests serve as dependent variables. Additionally, log data of the participants playing the personalized version of the analysis game serves as input to the exploratory analysis of in-game behavior.

Game Prototypes and Personalization

The primary intervention in this follow-up study is the personalized version of the analysis game (described in Chapter 6). In the approach to extend the current state of the art, the analysis game was adapted by utilizing the personalization pipeline described in Chapter 7. The objective was to provide personalized learning game content, and the new version of the game is referred to as *personalized game*.

Application of Personalization Pipeline

In the following, a brief overview is given on how the personalization pipeline was applied to the game: The personalization pipeline first provides a selection interface for learners to select services they either use, know, or do not know from a predefined set of services (e.g., 'PayPal', 'eBay'). The learners' selection is then used to compute a learner model (as described in Chapter 7.1.1). The learner model contains information about the learners' familiarity with different services. Next, URLs for all URL categories are created using a URL generator that applies different manipulation techniques to base URLs of a given service (e.g., manipulation of the registrable domain, subdomain, or path; for details, see Chapter 7). The learner model is used as input to the generator such that a set of URLs for different types of service can be created (i.e., services that learners use, know, or do not know). Generated URLs are then embedded into a level definition. The level definition is provided to the game to create a personalized version of the game for individual learners. The game purposefully includes some less well-known services to understand how participants handle such unknown services. The current version of the game presents known and unknown URLs at a 4:1 ratio. Due to randomness implemented in the rules used by the URL generator, returning learners will encounter different URLs compared to previous gameplay sessions. Beyond personalization, the game fully supports event logging of all in-game actions, timings, and results, which allows for the analysis of in-game behavior as part of this study (and in response to **RQ-3**).

In conclusion, the adaptation of the analysis game using the developed personalization pipeline results in the personalized game, a game prototype in which URLs are personalized based on the learners' familiarity with services.

Apparatus and Materials

Used questionnaires and tests

In order to compare the participant samples of both the original study and the follow-up study, the used survey with the different questionnaires were the same as before:

- **Perception of Phishing**: This questionnaire is an adapted subset of the questionnaire used in [Ara12] (see Chapter 8.1.2 for details on the adaptation). It measures the participants' perception of phishing as a risk using a 5-point Likert scale (from 1 = 'strongly disagree' to 5 = 'strongly agree') and was used in both pre- and post-test.
- **URL Classification Test**: This test consists of 20 (pre-test) and 30 (post-test) URLs that are to be classified as either benign or phishing URLs. Further, a question regarding the participants' confidence in their decision needs to be answered for each URL using a 6-point Likert scale. The test was included to measure effects on participants' performance and confidence, including the comparison of familiar and unknown URLs. Ten additional URLs were provided for the post-test to check for potential learning bias. Details regarding the test construction are described in Chapter 8.1.2. A complete list of all URLs used in the tests can be found in Table 41 in the Appendix.
- **Recognition of Services**: This questionnaire lists all services that were used to create URLs of the URL tests for participants to select for each service whether they (a) use it, (b) do not use it but know it, or (c) whether it is unknown to them. This test was included to analyze the effects of familiarity with a service on the participants' classification performance and confidence in the URL classification test. It was part of the post-test.
- **Demographics**: This questionnaire is used to collect demographic data of the participants, including age, gender, and educational background. It was used as part of the post-test.

Procedure

The procedure was not changed from the original study. The study was conducted as a remote lab study using video conferencing software and a web browser on participants' devices. It was structured into seven phases:

1. **Participant arrival**: First, participants joined the video conference session and were able to test their microphone, which would be needed in case of questions. Then, the instructors waited a sufficient amount of time until all participants joined, but no longer than 5 minutes after the initial start time of the session.
2. **Briefing**: As an introduction, participants received a briefing on the topic of the study (i.e., anti-phishing learning games) and the instructors presented a definition of phishing (see briefing instructions in Listing 1 in the Appendix).
3. **Pre-test**: Next, participants were presented with the pre-test part of the survey.
4. **Playing**: After finishing the pre-test, the survey software directed participants to either the analysis game or the personalized game in a second browser tab.

5. **Post-test**: After playing either one of the games, participants returned to the survey for the post-test.
6. **Debriefing**: When all participants finished the survey, a debriefing was used to inform the participants about the study's overall goal and answer open questions.
7. **Closing**: After answering all questions, the instructors thanked all participants and closed the session by terminating the video conference session.

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

Participants

The study was conducted with 89 participants. Since it was a follow-up study and to objective was to compare the analysis game and the personalized game, the data on the participant group playing the analysis game in the original study was used ($N_{\mathrm{A}} = 40$). The participant group playing the personalized game ($N_{\mathrm{P}} = 49$) was newly recruited but with similar methods, i.e., online by posting information about the study in different social network groups of universities and distributing it via university mailing lists. To keep the participant samples comparable, recruitment was, again, 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 CS. Due to the duration of the study, a financial incentive of 15 € was offered to each participant.

Therefore, the joint participant sample can be characterized as follows: 49 (55.06%) identified as female and 40 (44.94%) as male. The majority of participants were between 20 and 29 years old (76.40%), followed by age 30 or older (16.85%) and participants of the age 19 or younger (6.74%). Figure 19 shows the distribution of age and gender. The analysis of the participants' level of education revealed that most participants were students with either Bachelor's degrees or high school diplomas (82.02%). Other participants reported to have completed a Master's degree (15.73%) or vocational training (2.25%).

8.2.3 Results of the Pre- and Post-Test

In this section, the survey data is analyzed to answer the research questions defined in Chapter 8.2.1. Therefore, a series of analyses and statistical tests are used. For each test, two groups were considered depending on which game the participants played: the *analysis game group* and the *personalized game group*. The collected raw research data can be found in [D].

Measurements and notations

The number of correctly classified URLs divided by the total number of URLs was used for performance scores, i.e., a relative score. Similarly, the confidence level was computed as the participants' mean confidence of all classified URLs. Depending on the hypotheses used to answer a research question, different statistical tests were conducted with a significance level $\alpha = .05$. Here, parametric t-tests were used if no deviation from normality was detected in preliminary data screening. Otherwise, non-parametric testing was performed, e.g., Wilcoxon signed-rank test. Furthermore, an ANOVA was

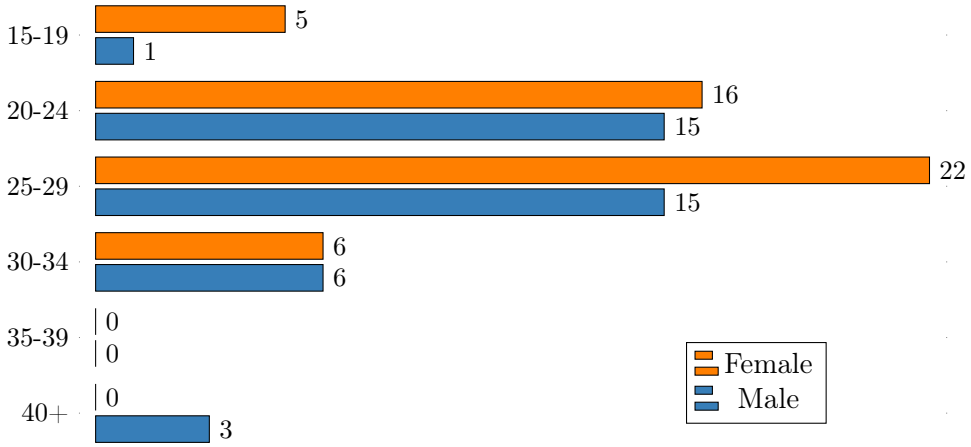

Figure 19: Distribution of Age and Gender ($N = 89$)

Table 21: Notations for statistical tests

Var.	Meaning
N	Sample size
M	Mean
SD	Standard deviation
α	Significance level
p	Significance value of statistical test
t	Test statistics for Student's (paired samples) or Welch's (independent samples) t-test
W	Test statistics for Wilcoxon signed-rank test
U	Test statistics for Mann Whitney U-test
F	Test statistics for ANOVA
ϵ	Greenhouse-Geisser estimate
d	Effect size by Cohen
r	Rank-biserial correlation coefficient
η_p^2	Partial η^2 estimates of effect size for ANOVA

used when testing differences between more than two groups. Similar to the first study, coherent notations for variables computed during statistical testing are listed in Table 21. Furthermore, the indices P and A are used for the personalized game group and the analysis game group.

Evaluation of potential learning bias

Before evaluating the research questions in detail, a check for a potential learning bias on URLs that were present in the pre-test is performed (see Table 22). By comparing $M_{\text{post-pre}}$ to $M_{\text{post-new}}$ using a one-tailed Student's t-tests with the underlying hypothesis that means for URLs that were also used in the pre-test are higher than the newly added URLs in the post-test. The test is not significant ($p > .725$), and the means are even higher for new URLs, concluding that a learning bias is negligible for the analyzed sample.

Table 22: Performance means in pre- and post-test including means on partial URL sets

Game	N	M_{pre} (SD)	$M_{\text{post-pre}}$ (SD)	M_{post} (SD)	$M_{\text{post-new}}$ (SD)
Analysis	40	.695 (.098)	.828 (.115)	.840 (.095)	.853 (.140)
Personalized	49	.726 (.114)	.811 (.110)	.823 (.104)	.855 (.123)

Differences between Pre- and Post-Test

Regarding the first research question (**RQ-1**), the following hypothesis was tested: The participants' performance in classifying URLs increased after playing either one of the games.

Statistical testing revealed that both games were generally effective: A Student's t-test for the analysis game group resulted in significant differences between the pre- and the post-test (as already presented in Chapter 8.1.3). Since for the personalized game group, a deviation from normality was detected (Shapiro-Wilk, $p = .033$), a Wilcoxon signed-rank test was used instead. Here, the results were significant as well, meaning that significant performance improvements occured in both game groups (see Table 23).

Table 23: Results of t-tests comparing relative scores in pre- and post-test for both games

Game	Test statistic	p-value	effect size
Analysis	$t(39) = -6.404$	$p < .001$	$d = -1.013$
Personalized*	$W = 775$	$p < .001$	$r = -0.717$

*Deviation from normality detected.

Differences between the Games

In response to **RQ-2**, the post-test results on all 30 post-test URLs for both game groups were compared, i.e., the analysis game group ($N_A = 40$) and the personalized game group ($N_P = 49$). The following hypothesis was derived: There are differences in performance and confidence between the two games.

Taking a look at the mean test results reveals that personalization did not lead to increased performances (see Table 22) or confidences (see Table 24). Even though the analysis game group performed better on average, this difference is not significant based on the results a two-tailed Welch's t-test ($t(85.891) = .797$, $p = .428$, $d = .169$, with no deviation from normality: Shapiro-Wilk, $p > .035$). Similar results could be observed for confidence levels: Here, the Shapiro-Wilk test was significant ($p < .001$) and a deviation from normality was detected. As such, a Mann-Whitney U-test was performed and returns no significant differences for the participants' confidence levels ($U(85.157) = 995.5$, $p = .901$, $r = .016$). This leads to inconclusive results and the rejection of the derived hypothesis.

Table 24: Confidence means in pre- and post-test including means on partial URL sets

Game	N	M_{pre} (SD)	$M_{\text{post-pre}}$ (SD)	M_{post} (SD)	$M_{\text{post-new}}$ (SD)
Analysis	40	4.065 (.637)	5.034 (.468)	5.086 (.461)	5.065 (.764)
Personalized	49	4.114 (.747)	4.948 (.655)	5.016 (.658)	5.259 (.478)

Effects of Personalization

Next, the effects of personalization are evaluated in response to the third research question (**RQ-3**). As such, similar to the first study, the participants' performance in classifying URLs of either unknown or known and used services is evaluated for the new participant sample which played the personalized game. For the analysis game group as well as the two other game prototypes tested in the first study, statistical testing revealed significant differences for URLs of unknown services and known services (see Chapter 8.1.3).

Since it might be possible that the personalization affected the classification results of different levels of familiarity in the tests, a repeated-measures ANOVA was performed comparing the three levels of familiarity, with the games as the between-group factor. It should be noted that the participant samples were adapted to $N_A = 34$ and $N_P = 39$, as some participants did not select any services as unknown, known, or used. Mauchly's test for sphericity is significant ($p < .001$), and Greenhouse-Geisser corrections are applied ($\epsilon = .728$). While no significant differences between the two games can be observed ($F(1, 71) = .084$, $p = .772$, $\eta_p^2 = .001$), there are significant differences between the levels of familiarity: $F(1.455, 103.308) = 10.204$, $p < .001$, $\eta_p^2 = .126$. Post-hoc tests (Holm) also show that unknown URLs are classified significantly less accurately than known and used in both games ($p <= .001$ in both cases), with no significant differences between known and used ($p = .525$). These results confirm the results of the initial study and show that there are differences in participants' performance when classifying URLs of services that are not familiar to the user.

While the study setup did not yield any significant differences in the participants' performance scores and confidence levels between the personalized game group and analysis game group, the issues with performance differences for URLs of unknown and known or used services raise the question of how learners interact with different URLs and how the personalization (i.e., familiarity with services) influences gameplay. For a closer look into possible effects of personalization, in-game log data was analyzed, and results are presented in Chapter 8.2.4.

8.2.4 Results of the Gameplay Analysis

In this section, the results of the gameplay analysis, i.e., an exploratory analysis of game log data generated in the personalized game, are described. The log data of the personalized game gives more insight into the learners' interactions with URLs of services that are used, known or unknown, during gameplay, as this information is not available for the original analysis game. The following analyses are all in response to the fourth research question described in Chapter 8.2.1.

Differences for Levels of Familiarity and Classification Outcomes

First, the learners' classification outcomes are analyzed with respect to different levels of service familiarity. Thus, mean values for the classification outcome of each level of service familiarity are computed per learner and then analyzed, e.g., in aggregation as the average classification outcome of all learners.

Table 25 shows the computed mean values for the participant sample. There are clear differences in the relative classification outcomes of learners, with URLs of unknown services being classified with the least accuracy (with a mean difference of .068 to URLs

Table 25: Mean values of in-game classification outcomes (SD).

Familiarity	N	Correct	Incorrect	Not classified	Time (sec)
Used	49	.680 (.170)	.186 (.108)	.133 (.117)	4.13 (1.39)
Known	49	.665 (.180)	.192 (.142)	.143 (.115)	4.11 (1.69)
Unknown	49	.597 (.221)	.250 (.187)	.154 (.187)	4.27 (1.62)

of known services and .083 to URLs of used services). The accuracy difference between URLs of known services and URLs of used services is less definite (i.e., only a difference of .015).

Differences for Levels of Familiarity and URL Categories

Next, the differences of correct classification outcomes per familiarity level per category are assessed to understand better which categories of URLs contribute to the overall difference in classification outcomes for different levels of service familiarity. As there are many comparisons if all possible levels of familiarity and categories are considered, the focus lies on the percentages of misclassifications (phishing URLs as benign or benign as phishing URLs), per familiarity level per URL category present in the game. Table 26 contains the mean values for the different types of misclassifications as well as the number of valid (and missing) values per category per familiarity, as some learners did not classify any URLs of a particular category and level of familiarity, e.g., URLs of the 'Path' category that learners were not familiar with (i.e., unknown service).

Table 26: Misclassification per type per familiarity

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

There are only minor differences between the familiarity levels for the URL categories 'Path', 'IP', and 'Random' (mean differences <= .02). The mean values for these categories were relatively low, indicating that they were generally detected well. URLs of

the categories 'RegDomain' (<= .137) and 'No-Phish' (<= .044) have notable differences, with the highest rates of mistakes for unknown services. This indicates potential problems with URLs of unknown services, as learners fail to classify them correctly. Interestingly, the classification accuracy for URLs of the 'Subdomain' category is highest for unknown services (<= .017). It should be noted that a large number of possible combinations of familiarity and categories may lead to a higher probability of these differences occurring by chance.

Differences for Levels of Familiarity and Learners' Discarding Behavior

Finally, differences in learners' discarding behavior are analyzed for the different levels of service familiarity. Here, the question arises whether there are differences in learners' actions to discard or not classify a URL if the respective service is unknown compared to other levels of familiarity (see Table 25) It seems that even though the differences are minor (> .012 mean difference), URLs of unknown services were either actively skipped or opened but not classified more often than known or unknown URLs. Similar results can be observed for the timing data, where learners seem to take longer for unknown URLs (i.e., a difference pf >= .14 seconds). However, since both differences are relatively small, a larger data sample would be interesting to allow for proper interpretation of the results.

In conclusion, the detailed analysis of gameplay available for the personalized game indicates that URLs of unknown services are classified with less accuracy and discarded more often within the game than URLs of other service familiarity levels, namely used or known services. The main difference in the classification outcomes seems to be evident in the URL categories 'RegDomain', 'Subdomain' and 'No-Phish'.

8.2.5 Discussion

In the previous section, the results of the follow-up user study and gameplay analysis were described in response to the research questions presented in Chapter 8.2.1. While the results show no significant differences in the participants' performance scores and confidence levels between the two games (**RQ-2**), there were significant improvements between the pre- and post-test (**RQ-1**) as well as significant differences between the levels of service familiarity, i.e., unknown URLs were classified significantly less accurately than known and used URLs (**RQ-3**). The results of the gameplay analysis revealed that the accuracy of classifying URLs of categories differs depending on the level of service familiarity (**RQ-4**).

In the following, issues and open questions regarding the overall study setup and the results of the user study are discussed. Further, possible limitations and future work directions are presented.

Study Setup

The chosen study setup uses a pre-/post-test between-group design comparing two versions of the analysis game (described in Chapter 6). It is similar to the previous study design, as this allows for the comparison with the previously recruited participant sample for the analysis game (see Chapter 8.1).

Similar limitations due to study setup

Based on the similarities to the initial study setup, similar limitations apply to the setup of the follow-up study. As such, the results' generalizability is limited as the participant sample now (for the personalized game) and then (for the analysis game) consists mainly of students in their twenties. Here, replication with a different demographic and larger sample size would be beneficial to strengthen the generalizability of findings. Beyond that, the study setup still fails to raise and test for situational awareness as the games do not convey the knowledge and awareness of how and when phishers lure users into disclosing personal information, and the chosen study setup as a remote lab study does not simulate a real-world setting. Lastly, since the same URL classification test is used as before (in order to compare both samples properly), limitations described in Chapter 8.1.4 apply to the follow-up study as well.

Evaluation in real-world settings and over time

For future work, improvements of the game (i.e., the analysis game in its personalized and non-personalized version), and the setup to evaluate the game's effectiveness should be considered. For one, the personalized version of the game could be extended towards more information on how and when phishers trick users. Based on this extension, it would be interesting to adapt the study setup and use simulated phishing attacks to see whether the personalized game group may perform better in more realistic scenarios and over a more extended period. This way, the effects of personalization could be evaluated not only in a longitudinal setup but also in a less artificial test context (i.e., the URL classification test). The first step in this direction is achieved with the longitudinal testing presented in Chapter 8.3. However, for a study using simulated phishing attacks to test the learners' awareness, an ethics committee should review the study setup first.

In all, the follow-up study and the initial study share similar limitations as their design was kept the same to compare participant samples. Future work directions include the possible reproduction of the study to strengthen and confirm findings presented here or a redesign to evaluate long-term effectiveness in a more realistic setting.

Study Results

As described in Sections 8.2.3 and 8.2.4, the results show significant performance improvements between pre- and post-test (**RQ-1**), but no significant differences for the participants' performance between the personalized and non-personalized version of the analysis game (**RQ-2**). Differences were also found for the participants' performance in classifying URLs of different familiarity levels with services (**RQ-3**).

Performance improvements between pre- and post-test

The results regarding the first research question (**RQ-1**) are in confirmation with the findings of the initial study, in which all games lead to improvements between the pre- and post-test. This is not surprising since the learning intervention, i.e., the game, is designed to convey the relevant knowledge and skills so that participants' can classify URLs in the URL classification task. To rule out potential learning bias on the URLs in the pre-test, additional URLs were added to the post-test, and rather high performance scores for these URLs confirm the general improvement of the participants' abilities (see Chapter 8.2.3).

Differences between the two games

The results of comparing the two versions of the analysis game did not yield any significant differences (**RQ-2**), meaning that there might be no difference between the effectiveness of the personalized and non-personalized version of the game. However, absence of evidence does not imply evidence of absence, such that a general effect of personalization should not be dismissed. It is merely not significant in the select participant sample, and hence, generalizability is again limited. Here, replication of the study with a larger but also more diverse participant sample could provide more insights

into the potential differences between the two versions of the game. On another note, the URLs in the analysis game were customized, i.e., selected based on the overall scope and expected participant sample of German-speaking adult end-users with regular online activity. As such, selected URLs in the test and game have a high chance of being known by the participants, limiting the power of personalization in the personalized game.

Controlling for service familiarity

The one difference between the two versions that is guaranteed is the inclusion of the selection interface in the personalized game and hence, the available knowledge about which services learners use, know, or do not know within the game. Since the results show significant differences for classifying URLs of unknown services compared to other levels of familiarity (**RQ-3**), the configuration of the personalized game could be an interesting manipulation parameter. In particular, it is possible that fixing the ratio of unknown services that appear in the game to different values (.2 in the current version) or integrating explicit instructions and game mechanics to deal with unknown URLs might have an impact on the learning outcome or awareness after playing the games. Thus, it can be argued that content personalization makes sense in the context of anti-phishing games, as it may offer a more immersive experience, in addition to the advantages when analyzing in-game data. Different personalization options and the effect on the learners' level of awareness might be worth exploring in more detail in future work. A particular direction is personalization through adaptivity, i.e., presenting personalized content adaptively depending on the learners' actions in the game. This way, learners could be supported in their learning experience by adapting the gameplay and tailoring the game even more.

Differences in gameplay

In response to **RQ-4**, the analysis of gameplay of the personalized game showed that there are some URL categories with a more significant difference in accuracy when classifying unknown URLs. Though it makes sense that the 'RegDomain' and 'No-Phish' categories have a high impact, as these URLs can be ambiguous if the original domain is unknown, another interesting is that the classification of URLs in the 'Subdomain' category was performed with higher accuracy for unknown URLs. Learners may parse the domain name from left to right and stop if they recognized a familiar domain name or saw a deceptive keyword but continued and correctly identified the actual registrable domain for unknown services. As the difference is small compared to the other categories, it is also possible that the difference is due to chance. A general problem with the gameplay analysis is that learners might have different strategies when playing the game, e.g., first opening a large number of coins and only classifying the easiest ones. These strategies might have affected the analysis outcomes. In particular, some differences might have been inflated by a small number of learners. As such, a more detailed analysis of learner strategies and their effect on the log data might be a possible addition in the future.

Summary of discussion and future work directions

In all, the results of the follow-up study show that personalization has interesting effects on the participants' classification abilities. Although the results did not yield significant differences between the personalized game and the analysis game, the participants' performance improved after playing either one of the games. Similar to the first study, differences in service familiarity and the participants' performance were found, and the gameplay data of the personalized game allowed analyzing even more than survey data. While the study setup and the current version of the games did not exhaust more methods for personalization, results indicate that content personalization, which has not been explored in much detail in other domains either, is a worthwhile pursuit. Future work opportunities include the redesign of the personalized game to support adaptive gameplay on a macro or micro level in which learners' actions guide the continuation of the game and selection of personalized game content (similar to the

work in [LR08] and [KA10]). For example, the learners' classification outcomes on used, known, or unknown URLs could be analyzed immediately in one level. The next level's content could then be adapted depending on learners' mistakes and potential misconceptions identified in the previous level. This way, training effects could be increased by reiterating over potentially problematic URL categories to give learners a chance to improve before advancing in the game. Further future work lies in reproducing the results of the follow-up study with larger and potentially more diverse participant samples to strengthen the evidence and generalizability. Alternatively to the chosen setup, personalized phishing attack simulation could be explored, similar to the work on simulated spear phishing attacks [Kum+08; Lin+19].

8.3 Evaluation of Long-term Effects in a Longitudinal Study

In this section, a longitudinal study as the extension to both the initial study (see Chapter 8.1) and the follow-up study (see Chapter 8.2) is presented. First, the research objectives and questions are introduced in Chapter 8.3.1, followed by a description of the study setup (see Chapter 8.3.2). Next, the results are presented in Chapter 8.3.3. Lastly, the presentation of the study concludes with the discussion of the results in Chapter 8.3.4. Parts of this study were previously published in [I] and [J]. Research data is publicly available in [D].

8.3.1 Research Objectives and Questions

In the following, the research objectives of the longitudinal study are introduced as well as the underlying research questions that are addressed by the study. While the initial study and the follow-up study followed a pre-/post-test between-group design with different experimental groups (i.e., the different games), the study setup allowed only to evaluate immediate differences between before and after playing either one of the games. To this end, no long-term effects were evaluated, and the question of whether the games have long-term effects on participants' ability to classify URLs was not evaluated. To evaluate long-term effects, in both studies, participants were asked for their permission to be contacted again for further studies.

Research objectives and questions

The research objective of the longitudinal study is to evaluate possible long-term effects of the games in the scope of the URL classification test and self-reported behavioral changes. The goal was to evaluate how the participants' performance and confidence in classifying URLs changed after a more extended period of time. Furthermore, the longitudinal study should give insights into potential behavior changes among the participants and evaluate their interest in game-based anti-phishing education. As such, the following research questions were formulated:

- **RQ-1**: How does the participants' performance in classifying URLs change for pre-, post- and longitudinal tests?
- **RQ-2**: How does the participants' behavior change after the study?

8.3.2 Study Setup

The study setup was designed as a longitudinal test at least three months after participation in the initial study or the follow-up study. Similar to the pre- and post-tests, the longitudinal test measures the participants' performance and confidence in classifying URLs (as the dependent variables). The games serve as independent variables, i.e., the *analysis game*, the *creation game* and the *decision game* from the initial study as well as the *personalized game* from the follow-up study.

Apparatus and Materials

To evaluate the participants' performance and confidence in classifying URLs as well as possible self-reported behavioral changes after completing the pre-/post-test part of the study, a survey containing the following questionnaires or tests was used:

- **URL Classification Test**: This test consists of 30 URLs classified as either benign or phishing URLs. Further, the test includes a question regarding the participants' confidence in their decision for each URL using a 6-point Likert scale. A complete description of the test construction and its use in the initial study is provided in Chapter 8.1.2. For the longitudinal testing, ten additional URLs were provided to check for potential learning bias.
- **Behavioral Change Questionnaire**: This questionnaire consists of nine items on the participants' behavior and attitude towards phishing after playing one of the game prototypes in the pre-/post-test part of the study (see Table 27). Dividing the items into four categories provides insights into the self-reported application of knowledge, interest in learning more about security using games, behavioral change, and the perception of phishing as a threat. The items are used with a 6-point Likert scale (from 1 = 'strongly disagree' to 6 = 'strongly agree').

Procedure

The longitudinal study was conducted three months after the pre-/post-test. As such, participants of the study in November 2020 were tested again in March 2021, while the study participants in May 2021 were tested again in September 2021. Participation in the longitudinal study required completing another survey without any scheduled video conferencing session. The participants did not play any of the games again. Therefore, no additional guidance was needed while completing the survey.

Participants

For participation in the longitudinal study, participants of the initial study and the follow-up study were asked to provide their email addresses at the end of the survey used in the post-test. After downloading the survey data of the first two studies, collected email addresses were immediately removed from the data set and stored separately without any identifier linking them to the participants' answers in the pre- and post-test. The months later, participants were contacted via email and invited to participate by completing the additional survey. As an incentive for participation, a lottery of 4×10 EUR was offered (for the longitudinal test in March 2021 and the test in September 2021).

Table 27: **Behavioral Change Questionnaire** used in longitudinal testing with the four constructs: Application (App), Interest (Int), Behavior Change (BC) and Perceived Threat (PT). A German translation can be found in Table 40 in the Appendix.

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

A total of $N = 82$ participants completed the survey of the longitudinal study resulting in a dropout rate of 54.95%. Among the participants were 34 males and 48 females. The distribution of age and gender shown in Figure 20. The participant sample splits into four groups depending on the game the participants played in the initial and follow-up study ($N_A = 17$, $N_C = 25$, $N_D = 21$, and $N_P = 19$). Unfortunately, this limits the evaluation of longitudinal effects and calls for reproduction with a larger participant sample.

8.3.3 Results

In this section, the results of the longitudinal study are presented in order to answer the research question described in Chapter 8.3.1. As such, a series of tests is performed similarly to the tests for the evaluation of pre- and post-test data in the previous studies. To evaluate long-term effects, performance scores and confidence levels of the longitudinal test are compared to the respective results of the previous studies. The different participant groups for each game are the analysis game group, creation game group, decision game group, and personalized game group. The collected raw research data can be found in [D].

Measurements and notations

Tests were conducted with a significance level $\alpha = .05$. Depending on the question or underlying data, either a one-tailed Welch's t-test, an ANOVA, or the non-parametric Wilcoxon signed-rank test were used (for the used notations, see Table 28).

Differences between the Pre-, Post- and Longitudinal Tests

In response to the first research (**RQ-1**), a comparative analysis of the participants' performance scores and confidence levels is performed, and the following hypotheses are

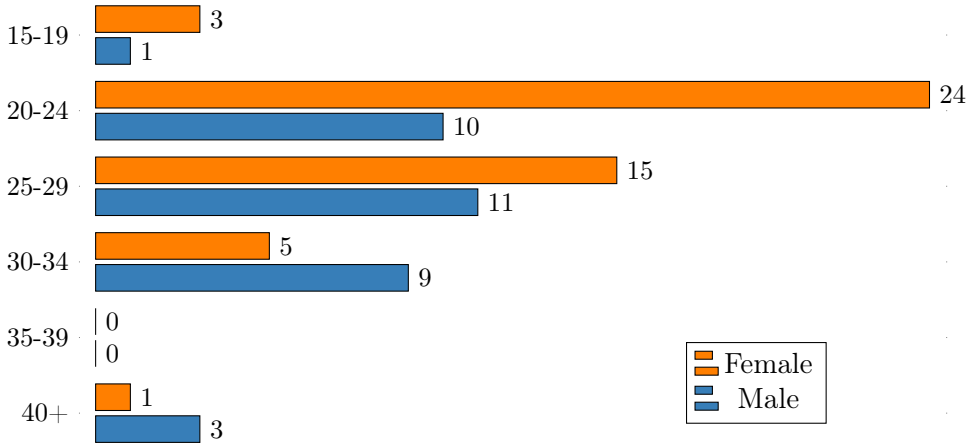

Figure 20: Distribution of Age and Gender ($N = 82$)

Table 28: Notations for statistical tests

Var.	Meaning
N	Sample size
M	Mean
SD	Standard deviation
α	Significance level
p	Significance value of statistical test
F	Test statistics for ANOVA
ϵ	Greenhouse-Geisser estimate
χ^2	Test statistic for Chi-square test
η_p^2	Partial η^2 estimates of effect size for ANOVA

tested: The participants' performance scores and confidence levels in the longitudinal test are better than the performance in the pre-test.

Evaluation of a potential learning bias

Table 29 presents the participants' mean performance scores of the URL classification test used in the pre-, post-, and longitudinal test (computed on the 20 URLs used in all tests) divided by the different game groups. Regarding a potential learning bias on URLs already used in the pre- and post-test, a comparison of mean values for the URLs in the pre- and longitudinal test ($M_{\text{long-pre}}$) and the newly added URLs ($M_{\text{long-new}}$) shows only small differences. Further, results show an even higher mean for new URLs for the personalized game group. As such, learning bias is considered to be negligible.

Table 29: Performance means in longitudinal test and comparison to pre- and post-test scores

Game	N	M_{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)
Creation	25	.698 (.130)	.762 (.105)	.754 (.131)	.740 (.119)
Decision	21	.686 (.107)	.821 (.092)	.774 (.119)	.771 (.096)
Personalized	19	.679 (.121)	.800 (.118)	.776 (.112)	.826 (.115)

Evaluation of performance scores

When evaluating descriptive statistics, the mean values of the participants' performance scores indicate a decline in performance between the post- and longitudinal test, with the pre-test performance score remaining the lowest for all four game groups. Even though the response rate was low and sample sizes of the longitudinal test were small, statistical tests were computed on the reduced samples of participants who completed all tests, i.e., the pre-, post-, and longitudinal tests. However, it should be noted that these tests are not as powerful due to the small sample sizes. Thus, even trends can be misleading. To test for the significance of the mean differences, a repeated-measures ANOVA was performed, using the three tests (pre, post, longitudinal) as repeated measures and the games as between-subject factors. Mauchly's test for sphericity was violated ($\chi^2(2) = 6.624$, $p = .036$). Thus, the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\epsilon = .98$). The results of the ANOVA confirm, that there are significant differences ($F(1.848, 144.121) = 49.555$, $p < .001$, $\eta_p^2 = .388$). Post-hoc tests with Holm correction show that pre-test performance is significantly lower than both post- and longitudinal test performances ($p < .001$ in both cases), but also the differences between post- and longitudinal tests are significant ($p = .004$).

Evaluation of confidence levels

Similar results to the improvements in the participants' performance can be observed for the reported confidence levels. Table 30 contain the participants' mean confidence values for the URLs used in all three tests as well as for the new URLs in the longitudinal test ($M_{\text{long-new}}$) and the overall mean confidence in the complete longitudinal test (M_{long}). Data exploration using descriptive statistics indicates overall improvement since the values in the longitudinal test are consistently higher than in the pre-test. Beyond that, the results of a repeated measures ANOVA also confirm significant differences between pre-, post- and longitudinal tests ($F(1.578, 123.060) = 39.841$, $p < .001$, $\eta_p^2 = .338$, after Greenhouse-Geisser corrections, $\epsilon = .789$). Additional post-hoc tests with Holm correction result in significant differences between the pre-test and both post- and longitudinal tests ($p < .001$ for both) and also between the post-test and the longitudinal test ($p = .018$). However, similar to the statistical testing on performances scores, the small sample size limits the interpretation of the results.

Table 30: Confidence means in longitudinal test and comparison to pre- and post-test

Game	N	M_{pre} (SD)	$M_{\text{post-pre}}$ (SD)	$M_{\text{long-pre}}$ (SD)	$M_{\text{long-new}}$ (SD)
Analysis	17	4.088 (.705)	4.968 (.497)	4.676 (.405)	4.629 (.535)
Creation	25	4.166 (.702)	4.720 (.649)	4.578 (.709)	4.764 (.703)
Decision	21	4.186 (.821)	4.993 (.478)	4.700 (.739)	4.676 (.728)
Personalized	19	4.318 (.609)	4.911 (.666)	4.850 (.692)	4.895 (.654)

Overall, the hypothesis can be accepted as the mean values show a trend of improved performance scores when comparing pre-test scores and longitudinal test scores. Statistical testing using a repeated-measures ANOVA confirm this, but the results should be interpreted carefully due to small participant samples.

Evaluation of Self-Reported Behavioral Changes

Regarding the second research question (**RQ-2**), the results of the Behavioral Change Questionnaire are analyzed using data exploration. As explained in Chapter 8.3.2, the behavioral change questionnaire consists of nine questions divided into four constructs. These constructs evaluate whether the participants have applied lessons from the game

after playing (Application), how interested the participants are in security-related learning games (Interest), whether the participants changed their everyday behavior after playing the games (Behavior Change), and to what extent the participants perceive phishing as a threat (Perceived Threat).

While the overall results are relatively positive (see Table 31), the answers to the questionnaire are self-reports by the participants, and for a self-reported measure, a certain amount of bias can be expected. Comparing the mean values with the setup of the follow-up study in mind, minor differences between the analysis game group and the personalized game group can be seen in all constructs, i.e., higher mean values for the personalized game group in all constructs. Meanwhile, the decision game group has the highest scores for the constructs Application and Behavioral Change. Due to the small sample sizes for each game and no clear trend in the descriptive statistics, no further statistical testing was performed.

Table 31: Results of Behavioral Change Questionnaire with constructs Application (App), Interest (Int), Behavior Change (BC) and Perceived Threat (PT).

Game	N	$M_{\text{App}}(SD)$	$M_{\text{Int}}(SD)$	$M_{\text{BC}}(SD)$	$M_{\text{PT}}(SD)$
Analysis	17	3.509 (1.285)	4.059 (0.966)	3.853 (1.412)	4.176 (0.557)
Creation	25	3.902 (1.383)	4.540 (0.803)	4.080 (1.320)	4.280 (1.225)
Decision	21	4.398 (0.827)	4.238 (1.032)	4.357 (0.761)	4.262 (0.785)
Personalized	19	4.071 (1.275)	4.684 (1.121)	4.105 (1.174)	4.368 (1.141)

8.3.4 Discussion

With the results of the longitudinal study presented in the previous section, a discussion of the study setup and the study results is described in this section.

Summary of findings

Regarding **RQ-1**, the analysis of the participants' performance scores and confidence levels shows significant differences between the pre-test and both the post- and longitudinal tests, which indicates possible knowledge retention for at least three months after playing either one of the games. For **RQ-2**, an exploratory approach revealed rather high scores for all constructs of the behavioral change questionnaire (see Table 31 for the results and Table 27 for the complete questionnaire). Furthermore, the comparison of the personalized game and the analysis game showed higher mean values in all four constructs for the personalized game.

In the following, a discussion of issues and open questions regarding the overall setup and results of the longitudinal study is described, and future work directions are outlined.

Study Setup

The longitudinal study consisted of another survey participants were asked to complete three months after the pre-/post-test study (either the initial study presented in Chapter 8.1 or the follow-up study presented in Chapter 8.2). This setup was chosen to evaluate the long-term effects of the games and further ask about changes in the participants' behavior and perception. As participation in the longitudinal part of the study was independent of the first part (i.e., pre- and post-test), not all participants decided to take part in the additional survey. This led to a high dropout rate of 54.95% and only

82 participants (compared to 182 participants in the first studies). The question arises whether only already-interested participants agreed to take part in the longitudinal test, which introduces additional self-selection bias to the results of the behavioral change questionnaire. Furthermore, results cannot be generalized due to the small participant sample. The reproduction of this study with a larger sample size is recommended to strengthen the evidence base. To better understand long-term effects, it would also be interesting to design the study with even more extended periods for longitudinal testing to see how the participants' performance scores and confidence levels change and whether regular repetitions might be needed for the knowledge to remain present.

Study Results

With the analysis of the longitudinal test, the results show that while the mean performance scores decreased compared to the post-test immediately after playing the game, the mean performance scores were still higher than the pre-test (**RQ-1**). Even though the sample size was small, the difference between pre- and longitudinal tests was significant, which implicates that the knowledge conveyed in the games was retained, at least partly, as this may only hold for the participants of the longitudinal test. Similar results were also observed for the participants' confidence levels. Due to the significant dropout rate and the possible self-selection bias, these results cannot be generalized, and more supporting evidence by replication studies would be needed.

Limitations of Behavioral Change questionnaire

The results of the behavioral change questionnaire present high scores in all four constructs for all four game groups and mean values for the personalized game group are higher than the analysis game group, which might indicate a more substantial influence on the participants' perception and behavioral change. However, it should be noted that these results rely on self-reported data from a custom questionnaire designed to be used in this study setup. Thus, these findings should only be seen as the first indicator of differences when including personalization, but they are far from conclusive evidence. Personalization may make the game more appealing and its learning content more transferable to the real-world contexts where users have to deal with potential phishing attacks from services they know and use. Future work might explore how simply making personalization options more present might already lead to a more immersive or relevant gaming experience. In addition, an evaluation using the questionnaire with a larger participant sample and after a review through domain experts is recommended to strengthen its quality and suitability for future studies.

Summary of discussion and future work

Overall, the longitudinal study serves as an additional evaluation of participants' test performance as well as self-reported changes in behavior and perception. The results are not generalizable due to a high dropout rate and, thus, a small sample size. In addition, results might include a certain amount of self-selection bias, as only interested participants decided to complete the additional longitudinal study survey. Similar to the previous studies, the effects on participants' situational awareness are difficult to assess and were not considered in the study setup. Replicating the results to strengthen the evidence and improve generalizability is recommended for future work. Further, a redesign of the study setup with less lab-like conditions for testing could be explored, e.g., through simulated phishing attacks over more extended periods. As already discussed for the initial and the follow-up study, the games' overall objective was to convey knowledge about the basic structure of URLs and possible manipulation techniques for phishing. Raising situational awareness and learning how and when phishers lure potential victims into disclosing personal information was not part of the

games. As such, future work directions may also include redesigning the games and improving the games' learning content.

9 Lessons learned from the Evaluation of Different Game Prototypes and Personalization

This chapter covers a summary of the research process and results, a discussion of the research design and a discussion of the lessons learned.

9.1 Summary of Research Process and Results

In the scope of this dissertation, mainly four parts defined the overall research process. In the following, each part is summarized, focusing on the results and their contribution to this dissertation.

Systematic literature reviews

First, the review of immediate related work on personalization of anti-phishing learning games showed that the domain had not been systematically explored yet, and no prior work on the personalization of learning content for anti-phishing education was found. Thus, learning games for security education and anti-phishing education were conducted using two systematic literature reviews (see Chapter 5):

- In the first literature review, learning games for security education were retrieved from scientific literature and through a product search. They were then analyzed regarding their intended target group, educational context, and actual availability. The underlying hypotheses stated that there were not many learning games targeted at end-users and that available games do not teach sustainable skills and knowledge to educate the target group to behave securely and assess risks appropriately. The results show that while 99 games were found at first, only 39 games were still available. Among these, only 20 games were suitable for the target group of end-users and did not require any prior knowledge in CS or IT security. In addition to analyzing the intended target group and educational context, the G/P/S model, a classification scheme for serious games, was used to classify existing games according to the dimensions: gameplay, purpose, and scope [DAJ11]. The classification using the G/P/S model confirms the results regarding intended target groups and markets. When analyzing the content of available games for end-users, results indicate that learning content is often limited and presented out of context. However, this may not apply to all games since some games designed for formal educational contexts that could also be suitable for end-users were found to be rich in learning content and provided meaningful contexts as well. Overall, the results of the content analysis were inconclusive, which might be because of the broad domain of security education, in which some topics might be presented in varying quality to different target groups. As such, a more narrow, topic-specific second review of existing learning games was proposed to allow a more thorough analysis of learning content and learning goals.

- In the second review, learning games for anti-phishing education, a particular topic of security education, were similarly retrieved from scientific literature and through a product search. The underlying hypotheses were divided into three different types of analysis: First, the intended target groups and educational contexts were analyzed to test whether the majority of available games were designed for end-users with little to no prior knowledge in IT security and for the use in informal learning contexts. Next, an analysis of learning goals and game mechanics was performed to evaluate if used game mechanics aim at specific, shared learning goals and if learners are required to utilize procedural knowledge in the games. Lastly, the learning content of the games was analyzed with the underlying hypotheses that games convey detailed conceptual knowledge but do not contain knowledge on advanced contemporary attacks. The analyses show that while anti-phishing games are primarily targeted at end-users in informal learning contexts, limited availability is an issue again. Developed anti-phishing learning games are no longer available and fail to reach their intended target group. When analyzing learning goals using BRT, the results show that most games focus on remembering and understanding factual and conceptual knowledge. According to the taxonomy, only a few games address other cognitive processes and procedural or meta-cognitive knowledge. Often games implement a binary decision mechanic in which learners have to classify a URL or email as legitimate or phishing. While this process mimics real-life actions, it provides only limited practice opportunities and hinders detailed assessment of the learners' gained knowledge and skills. Regarding the learning content analysis, results show that while some games are better than others in their level of detail, most games fail to provide detailed conceptual knowledge. Furthermore, the results show that games lack information on advanced contemporary phishing techniques and attacks. Overall, the review shows that most games are limited in their use of different game mechanics and do not address a wide range of cognitive processes and types of knowledge (according to BRT). Thus, the current state of the art could be extended through either a redesign and adaptation of existing games or a new, improved design using different game mechanics. Learning content could be improved by formulating learning goals covering the complete taxonomy and providing detailed information on essential topics and advanced contemporary attacks.

Game design

Next, two new game prototypes were designed and implemented with two major design goals, derived from the results of the systematic literature reviews and the objective to be personalized and evaluated later on (see Chapter 6). Firstly, the design of the new game prototypes should extend the binary decision game mechanic of classifying phishing URLs to provide better assessment and feedback. Secondly, the new game prototypes should address higher-order cognitive processes of applying, analyzing, evaluating, and creating using different mechanics. For the selection of learning content, the main requirement was to provide knowledge and skills that are robust against adversarial influence and simple to understand by end-users while retaining general applicability. As such, the focus was set on the basic URL structure and different manipulation techniques used to create phishing URLs. The design was generally guided by BRT, and learning goals were defined for the development of two new game prototypes. The defined learning goals cover all cognitive processes of the taxonomy, and the learning content ranges from terminology, concepts, and principles to subject-specific techniques and methods, like URL parsing and manipulation techniques.

New game prototypes

As a result of the game development process, two learning games were implemented: the analysis game 'All sorts of Phish' and the creation game 'A Phisher's Bag of Trick'.

Both games follow an alternating tutorial-level structure, introducing new learning content and providing immediate practice opportunities. While the analysis game primarily fulfills the first design goal and extends the binary decision mechanic to a more analytical classification mechanic in which learners have to sort URLs into one of six different URL categories. This way, learners are less likely to guess, and classification outcomes provide insights into the learners' decision processes and possible misconceptions. If learners cannot decide, they can also choose to discard a URL. The creation game presents an alternative to the common task of classifying URLs in games and requires learners to apply manipulation rules to create their own malicious URLs. This way, learners can actively apply the rules they learn and manipulate different parts of URLs, which may foster a deeper understanding of how to detect manipulations in URLs. For the development of both games, the MTLG framework was used, i.e., a game development framework based on the HTML5 Canvas element and native JavaScript. The games are provided in the browser and implement basic event logging functionalities to support gameplay analysis for evaluation. The open, modular structure supported by the framework and the flexibility of JavaScript provide suitable conditions to integrate personalization. To evaluate and compare the game prototypes in user studies, a third game prototype was developed to serve as a baseline. The decision game is an exact clone of the analysis game in which the extended classification mechanic was replaced with a binary decision mechanic similar to related work. This way, a comparison of different game mechanics is possible since learning content and game structure are kept the same.

Personalization Pipeline

Next, a framework for the personalization of anti-phishing learning games is needed in order to compare the personalized and non-personalized versions of the game prototypes in user studies. Thus, the concept of the personalization pipeline was developed and implemented (see Chapter 7). The concept consists of three distinct steps in which first, learner data is collected to generate a learner model. Data is collected either manually or automated, i.e., through a manual selection interface in which learners select their level of familiarity for a given list of services or by using a browser extension that analyzes the learners' browser history and retrieves the most frequently visited websites. In addition, a hybrid approach combining both manual and automated data collection is considered but further followed in this work. Collected data is processed and stored in a learner model, i.e., an abstract representation that contains information about learners' familiarity with different services. In the next step of the pipeline, different content generators are used to generate personalized learning game content based on the information available in the learner model and the requirements for the game, i.e., a URL generator to generate personalized URLs for the games developed in the scope of this work. In the final step of the personalization pipeline, content is embedded in the games to create a personalized version for individual learners. While the concept was described as three steps, the resulting implementation of the pipeline consists of three modules containing different components and a suitable interface between them to allow for appropriate data flow as intended by the concept. The resulting personalization pipeline is a modular, extensible framework to personalize different anti-phishing learning games by generating personalized learning game content. Depending on the game, the learner model would need to be extended to store other learner characteristics, and different content generators would be required to create suitable personalized content, e.g., personalized phishing emails.

Evaluation

To evaluate the different game prototypes, with or without personalization, two different comparative user studies were conducted. The studies were conducted as remote

online lab studies and followed a pre-test/post-test between-group design in which different game prototypes were compared with different participant groups.

- In the first user study (see Chapter 8.1), the different game prototypes were compared without any personalization, i.e., the analysis game, the creation game, and the baseline implementation called the decision game. The evaluation focused on comparing differences in the participants' performance and confidence when classifying URLs as part of a URL classification test. Comparisons were made between pre- and post-tests and between the different game groups. Besides the participants' performance and confidence, familiarity with services was evaluated by comparing the participants' performance on used, known, and unknown services. Lastly, the participants' perception of phishing as a threat was evaluated as an indicator for awareness. The results of the user study showed that while there were significant improvements in the participants' performance after playing either one of the games, there were no significant differences between the different game groups. Furthermore, the results showed that service familiarity influences the participants' performance: Participants were significantly better at classifying URLs of services they know or use compared to unknown services. Significant differences were found for the participants' confidence levels between the creation game group and the other game groups, i.e., participants who played the creation game were less confident than participants of the analysis game or decision game.
- In the second user study (see Chapter 8.2), the personalization pipeline was used to create a personalized version of the analysis game (referred to as the personalized game). The study's objective was to compare the analysis game and the personalized game. Similar to the first study, the participants' performance and confidence were compared between the pre- and post-test and between the game groups. Also, familiarity with services and the perception of phishing were evaluated similarly as before. Since the study setup was kept exactly the same and only the independent variable was changed, i.e., the game, it was possible to compare the participant sample for the analysis game from the first study with the newly recruited participant sample for the personalized game. In addition to the statistical evaluation of pre- and post-test data, game log data of the personalized game was used in an exploratory gameplay analysis to understand further how personalization effects in-game behavior. Similar to the results of the first study, there were significant performance and confidence improvements between pre- and post-test, but not between the different game groups. Also, the analysis concerning the familiarity with services showed significant differences in the participants' performance between unknown and known (or used) services. The evaluation of game log data in an exploratory gameplay analysis revealed differences for levels of familiarity and URL categories, i.e., URLs of the 'RegDomain' and 'NoPhish' category are more difficult to classify if the service of the URL is unknown by the participants.
- For the evaluation of long-term effects and self-reported behavioral changes, a longitudinal study was conducted with all four participants groups of the different game prototypes (see Chapter 8.3). As such, an additional survey was sent to the participants three months after their first participation. The survey contained another URL classification test to evaluate the participants' performance and confidence but also a questionnaire on self-reported behavioral changes (measured using the four constructs: application of knowledge, interest in learning

more about security using games, behavioral change, and the perception of phishing as a threat). While a significant dropout was observed, the results indicate possible knowledge retention for at least three months, since for all games, the participants' performance was still significantly higher than the pre-test. Furthermore, the results show high scores for all constructs of the behavioral change questionnaire with higher means for the personalized game group when compared with the analysis game group. This might be seen as an indicator that personalization positively influences participants' experience as it may make the game more appealing and its learning content more transferable to their personal real-world contexts.

Conclusion

Overall, the work in this dissertation was guided by the review of related work and existing learning games in the domains of security education and anti-phishing education. The identified research gaps called for the design and implementation of new game prototypes extending the current state of the art and providing suitable interfaces to integrate personalized learning game content. Since no prior work on the personalization of anti-phishing learning games was available, a concept of a personalization pipeline was developed and implemented as a modular, extensible framework. The personalization pipeline was then used to create a personalized version of one of the game prototypes. To evaluate the different prototypes, three user studies were conducted to evaluate the effectiveness of the games in aspects of the participants' performance and confidence when classifying URLs, the differences between personalized and non-personalized content, and the long-term effects on participants' performance but also behavioral changes.

9.2 Discussion of Research Design, Methods and Tools

This section discusses the research design applied in this dissertation, and the use of different methods and tools is reflected. The research process followed an applied research methodology based on a combination of design-based research and experimental hypothesis-testing research (see Chapter 4).

Problem identification and literature review

As typical for the design-based research process, it starts with the analysis of a practical problem. In the scope of this work, the practical problem was identified by reviewing existing learning games for security education and anti-phishing education (see Chapter 5). Compared to the review of related work, these reviews approach literature search by carefully selecting search terms and retrieving a large number of publications which are then filtered and analyzed in multiple steps. While in the educational domain, problems are often identified by observation and careful analysis of educational contexts (e.g., in the classroom), for the context of game development and to improve game design, the method of systematic literature review with different analysis steps was suitable when extended to a product search to include possible non-scientific implementations of learning games. However, limitations may apply since no explicit review methodology was used in both reviews. This may limit the quality of the results since the literature base could be missing important publications in the respective fields, or the non-systematic analysis steps excluded relevant work. While the reviews were both done carefully and with best efforts, the review work could be repeated using a specific framework or methodology (cf. [GB09; Par+15]). To this end, review results could then be compared, and if new games are identified, they could be added to the POG data set (see Appendix B.1). As an alternative to the problem identification through literature

review, existing games could have been evaluated in user studies to identify problems, e.g., with the used game mechanics.

Design and implementation of the games and the personalization pipeline

Next, design-based research is characterized through the application of design and development steps to implement solutions for the identified problem. To this end, the research process of this dissertation included the design and implementation of two new anti-phishing learning games and the personalization pipeline. For game design, BRT was used to properly formulate specific learning goals and guide the selection of learning content. The design of the game prototypes was kept simple and somewhat traditional. Since no other game design framework was used, the games' design might be limited. It could be extended in a redesign step using a designated game design framework. For the development of the personalization, no design guideline was identified; however, related work (e.g., [IB18]) was considered for the implementation. In the implementation process of both the game prototypes and the personalization, the MTLG framework was used, as it builds upon the HTML5 Canvas element and native JavaScript. This way, the games and the pipeline are executed within the browser environment on the client-side, which may offer advantages for future developments, e.g., adaptive, personalized gameplay using the learners' browser history.

Evaluation with user studies

For the evaluation of the different game prototypes and the application of the personalization pipeline to create a personalized version of a game, the research process followed the experimental hypothesis-testing research design. As such, user studies were conducted to evaluate different aspects of the games, e.g., effects on performance and confidence, differences in service familiarity. The evaluation was structured in two user studies: the initial user study of comparing different game prototypes without personalization (see Chapter 8.1) and the follow-up study comparing a game prototype with its personalized version (see Chapter 8.2). For each study, specific research questions and hypotheses were formulated to guide the quantitative analysis of the results. The general study setup followed a pre-test/post-test between-group design with additional longitudinal testing. While the different game prototypes served as independent variables, dependent variables were provided by participants' performance scores and confidence levels in the URL classification test (used in pre-, post-, and longitudinal tests). Different experimental conditions were created based on the different games, and participants were assigned randomly to one of the conditions (which means playing one of the games). The studies were conducted in a remote online lab study setup in which participants used their personal devices to complete the surveys and play one of the games.

Questionnaires, tests, and statistical analyses

Similar to the objectives of experimental hypothesis-testing research, the objective of the different studies in this dissertation was to evaluate prior formulated hypotheses and discuss the results in order to identify new hypotheses for following studies and related work. A survey with different questionnaires and tests was constructed to collect necessary measurements and data about different dependent variables (e.g., the participants' performance). As such, a URL classification test was designed to measure the participants' performance and confidence when classifying URLs as legitimate or phishing. This approach is also used in related work (e.g., [Can+15; She+07]) and provides valuable results on the participants' improvements after playing an anti-phishing learning game. Further questionnaires were added to provide information on the participants' familiarity with services and demographics. For the analysis of the questionnaire and test results, different statistical methods were used. While in the pre- and post-test, participant samples allowed hypothesis testing using different methods of inferential statistics, the participant samples in the longitudinal tests were smaller due to a significant dropout. Thus, testing for longitudinal effects was done carefully, and the results

were only interpreted as indicators rather than definite implications. In conclusion, the use of questionnaires and tests in combination with statistical methods worked well. Similar to related work, the complete survey was suitable for the chosen study design and did not cause any problems during or after the study.

Evaluation of personalization

For the evaluation of personalization, the personalization pipeline was integrated into the analysis game and compared to its non-personalized version (see Chapter 8.2). While in the previous study, it was found that there are differences between known and unknown services in the participants' performance in the URL classification test, the follow-up study allowed a closer look into how personalization can affect the participants and their gameplay. In addition to statistical evaluation of possible differences between the personalized game and the analysis game, a gameplay analysis on game log data was performed. The gameplay analysis made it possible to see how the learners interact with URLs of known and unknown services. Further, learners' mistakes could be analyzed, and it was possible to draw connections to the learners' familiarity with services behind incorrectly classified URLs. While the approach was not completely exhausted in this dissertation, it provided valuable insights and helped identify future work opportunities.

Conclusion

Overall, the study design worked out well, even as a remote online lab study using video conferencing software to supervise participants. By randomly dividing the participants into different experimental conditions, it was reasonably easy to compare the participant samples of different game prototypes and compute possible differences of dependent variables. This way, formulated hypotheses were tested, and the results were further discussed to derive new hypotheses and to identify future work directions. The used questionnaires and tests provided suitable data for statistical evaluation, and only due to significant dropout in the longitudinal test, the evaluation of long-term effects was limited. As an alternative to the chosen quantitative approach, a more qualitative study setup with smaller participant samples and a stronger focus on the learners' experience could have been used to evaluate the games and personalization. However, this would require different research questions and hypotheses guiding the study design. Further, a field study design could have been explored to test the games and personalization in a more realistic scenario and possibly evaluate more awareness-related aspects. For the evaluation of personalization, the gameplay analysis provided interesting insights into how learners interact with different URLs and how mistakes correlate with service familiarity. While the analysis performed so far was somewhat limited, further exploration of the game log data could provide even more insights into the learners' strategies and learning behavior after making mistakes.

9.3 Discussion of Results and Lessons learned

This section presents a discussion of the lessons learned throughout the research process. First, the lessons learned from the review of related work and the two systematic reviews on learning games for security education and anti-phishing education are discussed. Next, results of the development process of game prototypes and the personalization pipeline are revisited. Lastly, lessons learned from the evaluation process are discussed.

Findings in related work

In Chapter 3, related work on personalization was reviewed in the domain of anti-phishing learning games, but also the neighboring domains of security education, learning games, and learning in general. While the immediate search for personalization in anti-phishing learning games did not yield any findings, the review of personalization in neighboring domains shows some research activity with possible influence on

the work in this dissertation. The review work is always limited to the search quality and the selected search terms. Since personalization and adaptivity are often used interchangeably, the question arises whether there are more, possibly domain-specific synonyms for personalization used in other related work. This would result in a larger body of related work, which might present some work on personalization. Conclusively, the identified research gaps are limited to the review of personalization in anti-phishing learning games and the quality of search for related work.

Results of the systematic literature reviews

For a more thorough review of existing learning games, two systematic literature reviews are presented in Chapter 5. While the first review focused on learning games for security education, the second review was a refined review on anti-phishing learning games. For both reviews, results show that although various games are designed for end-users in informal learning contexts, most games fail to reach the intended target group as they are no longer available. This might be because of the discontinuation of research projects, technical maintenance issues, or even loss of interest by the developers and researchers. However, the review of existing games is strongly limited to those still available. Especially when analyzing gameplay and learning content, non-available games can only be analyzed based on their publications. Often, these publications do not provide the necessary details or complete descriptions needed in the review process. This limits the power of the reviews since the publications can only present a limited view on the game and its design. While in a publication, details about the design choices are explained, and the overall gameplay might be described, it might still be challenging to assess the gameplay. In the scope of this work, both the analysis of literature and the analysis through gameplay were utilized to review existing games. In conclusion, the results of the reviews are limited due to the problem of availability, and if more games were available to play, lessons learned could be enriched with results from gameplay sessions, similar to the presented analysis of learning content (see Chapter 5.3.4).

Decisions during game development

In the next step of this dissertation, two learning game prototypes were developed using the results from related work and the systematic literature reviews. The lessons learned from the review work emphasized that existing games mainly focus on factual and conceptual knowledge and the cognitive processes of remembering and understanding (according to BRT). As such, there is a lack of anti-phishing learning games conveying procedural and meta-cognitive knowledge while addressing higher-order cognitive processes. Furthermore, many games are limited in the use of different game mechanics as they rely on a binary decision mechanic in which learners have to classify a given URL or email as legitimate or phishing. This limits the games' potential for detailed assessment and feedback, as the binary decision does not give any insights into learners' decision processes. Based on these findings, the objectives for designing new game prototypes were to explore different game mechanics and address higher-order cognitive processes. Two game prototypes, the analysis game and the creation game, were developed to fulfill these objectives and compare different game mechanics used in the new games in the scope of a user study. As an alternative, the game design could have focused on designing one game prototype, fulfilling both design goals simultaneously. For example, this could have been done using multiple types of levels incorporating different game mechanics to cover a wide range of cognitive processes and to go beyond the binary decision mechanic seen in existing games. However, the decision was made to design two independent games and to compare their effectiveness and the use of different game mechanics in a user study (see Chapter 8.1). To compare with related work, a third prototype, the decision game, was created that implements the binary decision mechanic similar to existing games.

Lessons learned from evaluation

For the evaluation of the different game prototypes and the personalization, different user studies were conducted and presented in Chapter 8. While the results, limitations, and future work opportunities were previously discussed in the context of each study, some lessons learned and results should be discussed in the scope of the complete dissertation.

Study setup and participant recruitment

Regarding the general study setup and methods for participant recruitment, it should be noted that due to the pandemic, a remote online lab study setup was chosen in which participants completed the whole study using their personal device from home. Although this setup worked out well and no problems occurred during the different sessions, the participants were mostly unsupervised during the study. As such, it cannot be ruled out that participants used external sources when filling out the URL classification test or when playing the games. Only minimal supervision was provided since participants received a briefing and debriefing or could ask for help during the study (in the video conference call). For future work, it could be recommended to replicate this study in a face-to-face setting and to evaluate whether participants request additional support or whether results differ significantly, which could imply a problem with the study setup. Regarding the recruitment of participants, the study was advertised online by posting about the study in different social network groups of universities and distributing it via university mailing lists. The different participant samples (for both studies) primarily consist of students aged 20 to 30, which may limit the generalizability of the results. Meanwhile, the online setup allowed participants from all over the world to participate. Compared to the online setup, a face-to-face session would most probably limit the participant sample to participants living in the Aachen area, further limiting the generalizability.

Limited comparability of games

When evaluating the three games in a comparative user study, the results show that all games lead to increased performance scores and confidence levels in the post-test. For the comparison of the different games, the results show that the analysis game group was not significantly different compared to the decision game group, which raises the question of whether the game mechanics make no difference at all or whether the study setup was not suitable to test possible differences between the games. The URL classification test was possibly not suited to highlight differences between the games, and the evaluation could be repeated using a different measurement for participants' performance and confidence. Further, the results show that the creation game group did perform worse than the analysis game and the decision game groups. While these differences in the participants' performance and confidence could be attributed to the different game mechanics, it could also be because of differences between the covered learning content in the games. The creation game was not as similar to the analysis game as intended, which is why the comparability might be limited. As such, lessons learned from comparing the analysis game and the creation game should be interpreted carefully. It is recommended to reevaluate the games after fixing differences in the learning content and creating possible new levels in both games to ensure enough practice. This alignment step was completed in a current, updated implementation of the games. The current game prototypes were corrected, and new levels were added to the creation game. In addition to evaluating more similar games, a combined game prototype could be evaluated and compared both the opportunity to create manipulated URLs and learn to classify given URLs.

Effects of personalization

For the evaluation of personalized content in anti-phishing learning games, the developed personalization pipeline (see Chapter 7) was integrated into the analysis game. This way, a personalized version of the analysis game can be created for individual learners. Since the creation game presented various issues in the first user study and

offered only limited personalized possibilities, it was not considered in the follow-up study. In the scope of the follow-up study, the personalized game was compared to the non-personalized version of the analysis game. The objective was to evaluate how personalization affects the participants' performance and confidence in classifying URLs of different service familiarity. Furthermore, an exploratory gameplay analysis using game log data of the personalized game group was used to understand how learners interact with URLs of different familiarity levels in the game. Initially, results showed performance differences for URLs of known and unknown services, indicating that the learners' familiarity with services may influence the ability to detect phishing URLs. In the follow-up study, comparing the participants' performance in the URL classification test did not yield any significant differences between the analysis game group and the personalized game group. Similar to the first study, the results showed differences in the participants' performance for classifying URLs of different service familiarity. This raises the question of whether the personalization of the analysis game was significant enough to influence the participants' gameplay and abilities in the URL classification test. To further analyze if personalization influences the participants' ability to detect phishing URLs in the game, the results of the exploratory gameplay analysis show differences in the classification accuracy for different URL categories when classifying unknown URLs. This may indicate that personalization affects learners' strategies in dealing with URLs of different familiarity levels. In future work, adaptivity could be utilized to control for these differences and guide learners throughout the game, e.g., by providing more practice opportunities for URLs that were classified incorrectly.

Personalization vs. customization

Since the evaluation results did provide strong evidence that personalization improves the participants' abilities to detect phishing URLs, it can be questioned whether personalization is too much effort and customization would have similar effects. To this end, the personalization pipeline could be adapted to collect fewer data about individual learners and aim to customize the games' content, i.e., personalization for a group of learners (e.g., learners living in Germany) instead of individual learners. While this would result in a certain level of fuzziness regarding URLs of specific services and the learners' familiarity with services, it could still be sufficient to evaluate whether expected service familiarity correlates with the difficulty of classifying URLs. If so, the approach to personalize anti-phishing learning games could be simplified to provide customizable versions, in which learners can select which version they want to play.

Effects on awareness and behavioral changes

Regarding the games' effects on awareness and behavioral change, the evaluation conducted in this dissertation focused more on comparing different game prototypes and assessing the participants' performance and confidence. While the participants' perception of phishing as a threat was evaluated before and after playing one of the games, there was no realistic measure of awareness. Since the participants were placed in a somewhat artificial test scenario, in which they had to classify different URLs as part of the URL classification test, the participants' awareness of when and how phishers lure potential victims into disclosing information was not assessed. For longitudinal testing, a questionnaire on self-reported behavioral change was used; however, due to a significant dropout resulting in a small size, its results are merely indicators and can only be interpreted carefully. Since the focus of the games and the evaluation itself was on the knowledge about URLs and different manipulation techniques, it was not a primary objective to evaluate effects on the participants' awareness. To this end, it is recommended to use the games with additional resources to learn more holistically about phishing.

Ethical considerations

Further, ethical considerations should be discussed in terms of the data collection and learner modeling in the personalization pipeline and the design and evaluation of anti-phishing learning games. In the context of this work, data about the learners is collected to create a learner model and generate personalized content for anti-phishing learning games. While the selection interface gives learners complete control over what they want to share and provide as an input to learner modeling, the browser extension is considered way more invasive and may limit the learners' control over what to share and when. In its current implementation, the browser extension processes all data on the learners' devices without leaking any information to external parties. Also, the games themselves are executed entirely in the learners' browser with no need for any server instance if logging functionality is deactivated. For the evaluation, log data was needed for the gameplay analysis, and as such, the browser extension was not used in the user studies. Therefore, the browser extension should be evaluated first in a smaller scope to assess its acceptance among learners and ensure a necessary level of transparency and trust before using it in the personalization pipeline. Regarding the general evaluation of anti-phishing learning games and personalization, it should also be considered that playing a single game without any context may lead to a false sense of security or unintentionally increase the fear of phishing. If learners perform well in the game and start to believe that detecting phishing URLs is easy, they may quickly overestimate their knowledge and skills. This can lead to a false sense of security as they feel safe even if they do not know when and how phishers lure potential victims into disclosing information. Although the learners might be able to detect a phishing URL, they may not know about other phishing attacks or even other security risks with similar characteristics, e.g., malware infestation. On the other side, it needs to be considered what happens if learners perform terribly in the games and have difficulties understanding how to detect a phishing URL. To this end, adaptive gameplay could provide additional support when difficulties are detected. Furthermore, additional support in the form of references to additional learning resources could be provided or advice on whom to contact, e.g., IT department or public organizations. As such, it is vital to frame the use of the games with additional information and provide context, e.g., a briefing and debriefing in the context of the user studies. Overall, it should be noted that the brief discussion of ethical considerations could be explored further in future work.

Conclusion

In conclusion, there are various lessons learned throughout the different parts of the presented research process. While the review of related work and existing learning games is limited to the search quality and the different analysis steps taken to identify research gaps and shortcomings, game development may be improved by using a proper game design framework. However, as the game prototypes were foremost developed for the integration of a personalization pipeline and the comparative evaluation, they were developed rather practically and with a focus on the knowledge, they should teach. For the development of the personalization, a concept of three steps was implemented in a modular and extensible component-based architecture. The pipeline was kept simple, and while it was used in a specific context, it can be easily adapted to fit other domains. Lastly, the evaluation compared different game prototypes with or without personalization. While the results do not show that the personalized game outperforms the non-personalized games, different analyses show differences in the participants' performance with URLs of unknown services. A first gameplay analysis of the personalized game shows that learners may have difficulties classifying URLs of unknown services for some URL categories. In the end, the results and lessons learned from this work are derived from different steps and decisions from the entire research process. As this work

is the first in the domain of personalized anti-phishing learning games, it may provide many open questions and opportunities for future work.

10 Conclusion and Outlook

This chapter summarizes the work of this dissertation in a brief conclusion and presents the contributions and responses to the posed research questions presented in Chapter 1.2. Lastly, opportunities and directions for future work are outlined to present possible next steps and open questions in the domain of personalized anti-phishing learning games.

10.1 Conclusion

The work presented in this dissertation can be placed in the domain of personalization in game-based anti-phishing education. Based on substantial review work of related work and existing learning games for security education and anti-phishing education, problems and research gaps were identified. While no prior work on personalized anti-phishing learning games existed to this point, different approaches to personalization were used in the neighboring domains of security education, learning games, and learning in general. Meanwhile, the review of existing learning games presented several shortcomings and problems regarding the games' design and learning content, thus, leading to the design and implementation of new anti-phishing learning games in the scope of this work.

The new learning games were designed to extend the state of the art by using different game mechanics and addressing a wider range of cognitive processes and types of knowledge (according to BRT). The game 'All sorts of Phish' (or short: analysis game) requires learners to classify given URLs into one of many categories, instead of relying on the commonly found binary decision mechanic used in existing learning games. The game 'A Phisher's Bag of Tricks' changes the game mechanic completely and asks learners to create their own malicious URLs by applying different manipulation techniques commonly used in phishing attacks. Both games were designed using BRT to formulate reasonable learning goals and select suitable learning content.

In the next step, a personalization pipeline was developed to allow for dynamic, personalized content generation in order to personalize the new anti-phishing learning games. The pipeline consists of three steps and is implemented using a modular, extensible component-based architecture. Data collection can be done manually using a selection interface requiring learners to select services they use, know, or do not know, or automatically with a browser extension that analyzes their browser history to collect visited websites or services. As a result, a learner model is created, which will be used for personalized content generation. Different content generators can create game-specific, personalized content, i.e., URLs, in the scope of this work. Lastly, a level generator embeds created content into level definitions that are then provided to the games to provide a personalized version of the game.

For the evaluation of developed anti-phishing learning games and a personalized game using the personalization pipeline, two comparative user studies were conducted following a pre-test/post-test between-group design. To test different hypotheses, the participants' performance, confidence, familiarity with services, and perception of phishing as a threat were measured and compared. Furthermore, an exploratory gameplay analysis using log data of the personalized game provided insights into the participants' behavior with URLs of known and unknown services within the game. While results showed the general effectiveness of all games by comparing the pre- and post-test performance, the comparison between the games revealed issues with the creation game, as participants performed significantly worse than the analysis game. The comparison of the analysis game and its personalized version did not yield significant differences; however, results show differences in the participants' performance and confidence between unknown and known (or used) URLs.

To summarize, this work presents a first approach to personalize anti-phishing learning games by providing a personalization pipeline integrated into two new game prototypes. Further, it presents an evaluation comparing the different game prototypes with each other as well as a baseline implementation and a personalized version to gain insights into how different game mechanics and personalized learning content may influence and support learners in detecting phishing URLs. This work contributes to the not yet systematically explored research domain of personalized anti-phishing learning games and may help find new research questions and directions for future work.

10.2 Contributions and Responses to the Research Questions

This section presents the main contributions made throughout the research process of this dissertation and draws connections to the research questions and objectives formulated in Chapter 1.2. As such, each SRQ will be answered by briefly summarizing what work has been done and which results have been achieved.

Sub research question 1

Regarding the first sub research question (**SRQ-1**), related work and existing learning games for security education and anti-phishing education have been reviewed in Chapters 3 and 5. The objective was to evaluate if personalization has been considered in the context of anti-phishing learning games and how the design of existing games could support individual learners using personalization. Since no prior work on the personalization of anti-phishing learning was found, neighboring domains were explored to evaluate how personalization is applied in security education, learning games or learning in general. The review of related work shows that mostly conceptual work or implementations realizing personalization by considering learning styles exist for security education. In learning games, personalization is most often achieved through different types of adaptivity. An alternative presents content personalization through procedural content generation supported by learner modeling. However, none of these approaches have been explored for game-based anti-phishing education.

Sub research question 2

In response to the second sub research question (**SRQ-2**), two new game prototypes were developed to fulfill two primary design goals derived from previous review work (see Chapter 6). Two systematic literature reviews were conducted analyzing the design and content of existing games to prepare for the development of new personalized anti-phishing learning games. The results show that many games are limited in their

use of different game mechanics and often implement a binary decision mechanic requiring learners to classify URLs or emails as legitimate or phishing. Furthermore, the games mainly focus on conveying factual and conceptual knowledge to remember and understand, thus, not addressing higher-order cognitive processes and other types of knowledge (according to BRT). Lastly, although games for end-users with little to no prior knowledge in IT security were developed in the past, only a few games are still available today, which limits their effectiveness as most games fail to reach their target groups. With two design goals derived from the results of the review work, two new learning games prototypes were developed in the scope of this dissertation. Both game prototypes implement different or extended game mechanics and address higher-order cognitive processes and different types of knowledge. Thus, more detailed assessment and feedback are possible to support learners in learning about the basic structure of the URL and different manipulation techniques used to detect or create phishing URLs.

Sub research question 3

For the third sub research question (**SRQ-3**), a framework for the personalization of anti-phishing learning games was designed and implemented (see Chapter 7). The framework is structured as a three-step personalization pipeline: First, a data collection module provides different methods to collect relevant data about the learners to compute a learner model as an abstract representation of learner characteristics (e.g., familiarity with services). Next, the learner model is provided as input to a content generation module in which different content generators are used to generate game-specific, personalized content (e.g., URLs). In the last step, the content delivery module provides a level generator component to embed generated content into level definitions. A level controller then delivers the personalized level definitions to the game, such that game-specific views and logic can be configured accordingly. Whenever learners advance through the game and new levels are needed, the level controller can trigger the generation of new levels and content. The modular, extensible architecture of the personalization pipeline allows for future use with different anti-phishing learning games and could also be extended to be used in different personalization contexts.

Sub research question 4

Regarding the fourth sub research question (**SRQ-4**), two user studies and an evaluation of long-term effects and self-reported behavioral changes are presented in Chapter 8. The objectives of the user studies were to compare the different game prototypes, with or without personalization, and gain insights on how they affect learners' performance in detecting phishing URLs. While in a first initial study with 133 participants, the different game prototypes were compared in a pre-test/post-test between-group design, a follow-up study with 49 participants allowed the comparison to a personalized version of one of the games. The results of the user studies were inconclusive, as the personalized game did not outperform the non-personalized games, and the new game mechanics did not lead to better performance scores than the baseline implementation. This might be due to the study setup, the used questionnaires and tests, or because there might actually be no differences between the game variants. Further, the analysis of service familiarity revealed problems with classifying URLs of unknown services, which confirms the motivation of this dissertation. For the follow-up study, an exploratory gameplay analysis using log data of the personalized game shows differences in the classification accuracy of unknown services for different URL categories. Regarding the self-reported behavioral changes, results show higher scores for the personalized game group compared with the non-personalized game group. This might indicate a more substantial influence on the participants' perception and behavioral change. However, due to a significant dropout, the results of the longitudinal testing are not generalizable and should be interpreted carefully. Overall, the evaluation provided various insights into how the different game prototypes may influence the participants' abilities to detect

phishing URLs and whether controlling for service familiarity using personalization can help to understand better how participants deal with different URLs. As the results are somewhat inconclusive, further evaluation is needed.

Main research question

In response to the previously stated main research question (**MRQ**) of how personalized anti-phishing learning games can be utilized to support end-users with little to no prior knowledge of IT security in detecting phishing URLs, four SRQs were formulated and addressed throughout this dissertation. As such, substantial review work regarding related work and two systematic literature reviews were conducted to identify research gaps and derive design goals for developing two new anti-phishing learning games. Further, a personalization pipeline was designed and implemented to be applied to the anti-phishing learning games. In the scope of two user studies and a longitudinal evaluation, different learning games, with or without personalized, were compared, and it was analyzed how personalization influences learners' abilities to detect phishing URLs.

Contributions

With completing the research process and addressing the posed research questions in this dissertation, four main research contributions can be identified for the research domain of personalization in game-based anti-phishing education:

- A systematic overview on existing literature and available implementations of learning games for anti-phishing education and security education (see Chapter 5)
- Two new anti-phishing learning game prototypes which use different game mechanics than existing learning games and address a wider range of cognitive processes and types of knowledge, but also allow more insights into the learners' actions and decision processes and provide suitable interfaces for the personalization of learning game content (see Chapter 6)
- A modular, extensible framework for the personalization of anti-phishing learning games using different modules for data collection and learner modeling, content generation, and content delivery (see Chapter 7)
- An comparative evaluation of the different game prototypes, with or without personalization, and insights regarding the influence of personalized learning game content on learners' performance in detecting phishing URLs (see Chapter 8)

While this research consisted of different steps and processes to design, implement, evaluate and understand personalized anti-phishing learning games, it is not finished and only presents the first work in the field of personalization of game-based anti-phishing education. It aimed to address different questions, extend the state of the art, and provide future work opportunities to gain more insights into how personalization can support learners in detecting phishing URLs.

10.3 Outlook

This section presents an outlook and identifies problems and challenges for future work. It introduces directions and opportunities for future research and development projects based on the findings and results presented in this dissertation.

Future work in the evaluation

In the scope of this work, the evaluation focused on comparing different game prototypes, with or without personalization. The user studies could be replicated with larger, more diverse participant samples to strengthen the evidence and provide more data on the different aspects evaluated so far. As such, recruitment strategies could be extended to find more participants. This could lead to more generalizability of results and new

hypotheses for future work. In contrast, the target group in user studies could also be specified more strictly, e.g., by focusing on younger teenagers and children between the age of 10 to 13 [R]. Furthermore, future work could adapt the study design by providing other instruments to measure the participants' performance and confidence when classifying URLs. Since the games mainly teach knowledge about the URL structure and manipulation techniques used for phishing, the evaluation focused on evaluating the participants' classification abilities and knowledge retention. To evaluate long-term effects and possible behavioral changes, more tests could be introduced to check the knowledge retention over extended time periods. Alternatively, the study setup could be changed from a lab study to a field study using simulated phishing attacks after participants played the games to see who still falls for phishing. This way, the game's influence on the participants' awareness could be evaluated as well.

For the evaluation of personalization, the exploratory gameplay analysis using methods of game learning analytics could be developed further. The analysis could be extended to understand learners' strategies when dealing with URLs of unknown services compared to known (or used) services. Furthermore, real-time analytics could be integrated into the games to support the implementation of adaptivity and to control the gameplay. A dashboard could be implemented to support instructors in a user study and provide insights into gameplay during the study, visualizing different indicators retrieved from analyzing the log data.

Future work in game development

Regarding the development of anti-phishing learning games, future work can go into different directions. Since no proper design framework was used in the development of the presented game prototypes in the scope of this work, future work could entail taking a step back to reevaluate their design and improve it using a specific design framework for learning games. Further, the analysis game and the creation game could be combined in a joint version requiring learners to analyze and classify URLs and apply manipulation techniques to create their own URLs. Both game mechanics could be utilized in the same game to provide more variety between levels, address more cognitive processes and foster different skills. Also, the learning content and gameplay could be adapted to support different strategies in dealing with URLs, e.g., for unknown URLs, learners should be taught to discard them immediately as this would mimic reasonable real-world behavior.

Since the developed games focus primarily on phishing URLs, different topics in anti-phishing education could be addressed by developing new learning games, e.g., a game about phishing emails. As such, an interactive email interface was developed, which requires learners to classify emails and mark deceptive attributes in suspicious emails [A]. Here again, learners could learn to detect phishing emails or even create their own phishing emails using different deception and persuasion techniques (e.g., [Cia06]). Two supervised student thesis projects show prototypical implementations of an anti-phishing learning game about phishing emails (see [1] and [4]). The first game prototype developed from these thesis projects is was evaluated and described in [K]. Besides expanding on other topics in anti-phishing education, new learning games could also address related topics of security education, e.g., certificates [2]. While in this work, the target group was selected very broad and without many restrictions, game design could also be tailored to a more specific audience. As such, an anti-phishing learning games for children and teenager between the age of 10 to 13 was developed (cf. [L], [9]). First evaluation results show similar effects on learners' performance and confidence, confirming findings of this work for a more specific target group (see [R] and [10]).

Future work with the personalization pipeline

Another future work direction in the scope of this dissertation lies in further exploration of personalization and improvements of the personalization pipeline. As such, new data collectors components could be implemented to retrieve other learner characteristics, e.g., by analyzing email content or browsing behavior over time. Since the browser extension has not been evaluated in its current form, future work could focus on creating a suitable evaluation scenario to see how automated data collection can be used to create appropriate learner models while preserving learners' privacy. Furthermore, hybrid data collection methods could be explored by combining the automated data collection with a browser extension and the manual selection interface to refine the learner model and make learner modeling more efficient. If in the future, the objective is to personalize learning games on email phishing or other topics, new content generators would be needed, e.g., an email generator to provide personalized emails. To this end, the personalization pipeline could also be applied to learning games in other domains, e.g., games for security education or language learning. In its current implementation, the pipeline may focus on anti-phishing learning games, but the modular and extensible design allows for the adaptation to other domains as well.

While only the analysis game was personalized and evaluated in a user study in this work, future work could explore how the creation game could be personalized. So far, the potential for personalization in the creation game may be limited due to the game's design, but possibly, by redesigning the game, it could be personalized and then evaluated in a comparative study setup with its non-personalized version. Furthermore, the architecture of the personalization pipeline provides a cyclic relation for on-demand content generation. This relation could be used to implement adaptivity in the games by including game-specific information during content and level generation, e.g., learners' mistakes could be analyzed and used to generate levels that provide more practice opportunities before learners advance in the game. To this end, a student thesis project explored how dynamic difficulty adjustment could be applied to the learning games in the scope of this work [12].

Conclusion

Overall, future work can extend into various directions, whether it is in support of findings presented in this dissertation or to explore further how content personalization can be utilized in anti-phishing learning games or related domains. Since this work only presents first approaches to personalize anti-phishing learning games, more work is to be expected to understand how personalization can be used to support learners in detecting phishing attacks.

A Bibliography

[AAC15] M. Alsharnouby, F. Alaca, and S. Chiasson. "Why Phishing Still Works: User Strategies for Combating Phishing Attacks". In: *International Journal of Human-Computer Studies* 82 (2015), pp. 69–82. DOI: 10.1016/j.ijhcs.2015.05.005 (pp. 95, 105).

[AAH15] M. Alshammari, R. Anane, and R. J. Hendley. "The Impact of Learning Style Adaptivity in Teaching Computer Security". In: *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*. Conference on Innovation and Technology in Computer Science Education. ITiCSE '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 135–140. DOI: 10.1145/2729094.2742614 (pp. 24–26, 31).

[AAS16] M. M. Ariffin, W. F. W. Ahmad, and S. Sulaiman. "Investigating the Educational Effectiveness of Gamebased Learning for IT Education". In: *2016 3rd International Conference on Computer and Information Sciences*. International Conference on Computer and Information Sciences. ICCOINS '16. Kuala Lumpur, Malaysia: IEEE, 2016, pp. 570–573. DOI: 10.1109/ICCOINS.2016.7783278 (p. 49).

[ABP18] D. Aladawy, K. Beckers, and S. Pape. "PERSUADED: Fighting Social Engineering Attacks with a Serious Game". In: *Trust, Privacy and Security in Digital Business*. International Conference on Trust and Privacy in Digital Business. Cham: Springer International Publishing, 2018, pp. 103–118. DOI: 10.1007/978-3-319-98385-1_8 (pp. 169, 170).

[ABS17] M. M. Al-Daeef, N. Basir, and M. M. Saudi. "Security Awareness Training: A Review". In: *Proceedings of the World Congress on Engineering 2017*. World Congress on Engineering. Vol. I. London, United Kingdom: Newswood Limited, 2017, p. 542 (p. 23).

[Abt70] C. C. Abt. *Serious Games*. New York, NY, USA: Viking Press, 1970. 176 pp. (p. 14).

[Agt+15] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis. "Seven Months' Worth of Mistakes: A Longitudinal Study of Typosquatting Abuse". In: *Proceedings 2015 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium. NDSS '15. San Diego, CA: Internet Society, 2015, pp. 1–13. DOI: 10.14722/ndss.2015.23058 (pp. 13, 65, 82, 88).

[AKA01] L. W. Anderson, D. R. Krathwohl, and P. Airasian. "A Revision of Bloom's Taxonomy of Educational Objectives". In: *A Taxonomy for Learning, Teaching and Assessing. Longman, New York* (2001) (p. 35).

[AL13] N. A. G. Arachchilage and S. Love. "A Game Design Framework for Avoiding Phishing Attacks". In: *Computers in Human Behavior* 29.3 (2013), pp. 706–714. DOI: 0.1016/j.chb.2012.12.018 (pp. 1, 11).

[ALK14] E. Amankwa, M. Loock, and E. Kritzinger. "A Conceptual Analysis of Information Security Education, Information Security Training and Information Security Awareness Definitions". In: *2014 9th International Conference for Internet Technology and Secured Transactions (ICITST)*. International Conference for Internet Technology and Secured Transactions. ICITST '14. London, UK: IEEE, 2014, pp. 248–252. DOI: 10.1109/ICITST.2014.7038814 (pp. 7–9).

[ALM15] N. A. G. Arachchilage, S. Love, and C. Maple. "Can a Mobile Game Teach Computer Users to Thwart Phishing Attacks?" In: *Journal of Infonomics* 6.3/4 (2015), pp. 720–730. DOI: 10.20533/iji.1742.4712.2013.0083 (pp. 2, 77, 169, 170).

[Alo+16] F. Alotaibi, S. Furnell, I. Stengel, and M. Papadaki. "A Review of Using Gaming Technology for Cyber-Security Awareness". In: *International Journal for Information Security Research* 6.2 (2016), pp. 660–666. DOI: 10.20533/ijisr.2042.4639.2016.0076 (p. 39).

[Amea] American Psychological Association. *Between-Subjects Design*. In: *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/between-subjects-design (accessed on 04.01.2022) (p. 36).

[Ameb] American Psychological Association. *Descriptive Statistics*. In: *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/descriptive-statistics (accessed on 04.01.2022) (p. 38).

[Amec] American Psychological Association. *Experimental Research*. In: *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/experimental-research (accessed on 04.01.2022) (p. 34).

[Amed] American Psychological Association. *Inferential Statistics*. In: *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/inferential-statistics (accessed on 04.01.2022) (p. 38).

[Amee] American Psychological Association. *Laboratory Research*. In: *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/laboratory-research (accessed on 04.01.2022) (p. 37).

[Amef] American Psychological Association. *Pretest–Posttest Design*. In: *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/pretest-posttest-design (accessed on 04.01.2022) (p. 36).

[Ameg] American Psychological Association. *Questionnaire*. In: *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/questionnaire (accessed on 04.01.2022) (p. 37).

[Ameh] American Psychological Association. *Statistical Analysis*. In: *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/statistical-analysis (accessed on 04.01.2022) (p. 38).

[APW21] APWG. *APWG Phishing Activity Trends Report, 1st Quarter 2021*. Anti-Phishing Working Group, 2021. URL: https://docs.apwg.org/reports/apwg_trends_report_q1_2021.pdf (accessed on 04.01.2022) (pp. 1, 11, 94).

[Ara12] N. A. G. Arachchilage. "Security Awareness of Computer Users: A Game Based Learning Approach". PhD thesis. Uxbridge: Brunel University, 2012. 237 pp. URL: http://bura.brunel.ac.uk/handle/2438/7620 (accessed on 11.06.2021) (pp. 12, 37, 96, 108, 112).

[Arn+15] S. Arnab et al. “Mapping Learning and Game Mechanics for Serious Games Analysis”. In: *British Journal of Educational Technology* 46.2 (2015), pp. 391–411. DOI: 10.1111/bjet.12113 (p. 51).

[AS18] H. Aldawood and G. Skinner. “Educating and Raising Awareness on Cyber Security Social Engineering: A Literature Review”. In: *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. International Conference on Teaching, Assessment, and Learning for Engineering. TALE’18. Wollongong, NSW, Australia: IEEE, 2018, pp. 62–68. DOI: 10.1109/TALE.2018.8615162 (p. 23).

[AVW20] S. Albakry, K. Vaniea, and M. K. Wolters. “What Is This URL’s Destination? Empirical Evaluation of Users’ URL Reading”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Conference on Human Factors in Computing Systems. CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–12. DOI: 10.1145/3313831.3376168 (pp. 12, 13).

[AZ17] A. Aleroud and L. Zhou. “Phishing Environments, Techniques, and Countermeasures: A Survey”. In: *Computers & Security* 68 (2017), pp. 160–196. DOI: 10.1016/j.cose.2017.04.006 (p. 1).

[BA19] G. Baral and N. A. G. Arachchilage. “Building Confidence not to be Phished Through a Gamified Approach: Conceptualising User’s Self-Efficacy in Phishing Threat Avoidance Behaviour”. In: *2019 Cybersecurity and Cyberforensics Conference (CCC)*. Cybersecurity and Cyberforensics Conference. CCC ’19. Melbourne, Australia: IEEE, 2019, pp. 102–110. DOI: 10.1109/CCC.2019.000-1 (pp. 169, 170).

[Bau+17] G. Bauer, D. Martinek, S. Kriglstein, G. Wallner, and R. Wölfle. “Digital Game-Based Learning with "Internet Hero": A Game about the Internet for Children Aged 9–12 Years”. In: *Context Matters!* Ed. by K. Mitgutsch, S. Huber, M. Wagner, J. Wimmer, and H. Rosenstingl. Wien: New Academic Press, 2017, pp. 148–161 (pp. 169, 170).

[BC14] C. Bergmann and G. Canova. “Design, Implementation and Evaluation of an Anti-Phishing Education App”. MA thesis. Darmstadt: Technische Universität Darmstadt, 2014. 113 pp. URL: http://tuprints.ulb.tu-darmstadt.de/3763/ (pp. 2, 58, 77, 169, 170).

[BC16] M. Baslyman and S. Chiasson. “"Smells Phishy?": An educational game about online phishing scams”. In: *2016 APWG Symposium on Electronic Crime Research*. APWG Symposium on Electronic Crime Research. eCrime ’16. Toronto, Ontario, Canada: IEEE, 2016, pp. 1–11. DOI: 10.1109/ECRIME.2016.7487946 (pp. 169, 170).

[Ber11] Z. L. Berge. “If You Think Socialisation in mLearning Is Difficult, Try Personalisation”. In: *International Journal of Mobile Learning and Organisation* 5.3/4 (2011), p. 231. DOI: 10.1504/IJMLO.2011.045314 (p. 30).

[Bha19] J. Bhardwaj. “Design of a Game for Cybersecurity Awareness”. MA thesis. Fargo, ND, USA: North Dakota State University, 2019. 45 pp. URL: https://hdl.handle.net/10365/29758 (pp. 169, 170).

[Blo56] B. S. Bloom. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. London: Longman Group, 1956 (p. 35).

[Boy+16] E. A. Boyle et al. "An Update to the Systematic Literature Review of Empirical Evidence of the Impacts and Outcomes of Computer Games and Serious Games". In: *Computers & Education* 94 (2016), pp. 178–192. DOI: 10.1016/j.compedu.2015.11.003 (p. 15).

[BP16] K. Beckers and S. Pape. "A Serious Game for Eliciting Social Engineering Security Requirements". In: *2016 IEEE 24th International Requirements Engineering Conference*. International Requirements Engineering Conference. RE '16. Beijing, China: IEEE, 2016, pp. 16–25. DOI: 10.1109/RE.2016.39 (pp. 169, 170).

[BPF16] K. Beckers, S. Pape, and V. Fries. "HATCH: Hack and Trick Capricious Humans - a Serious Game on Social Engineering". In: *Proceedings of the 30th International BCS Human Computer Interaction Conference: Companion Volume*. HCI '16. Swindon, GBR: BCS Learning & Development Ltd., 2016, pp. 1–3 (pp. 169, 170).

[BS04] S. Barab and K. Squire. "Design-Based Research: Putting a Stake in the Ground". In: *Journal of the Learning Sciences* 13.1 (2004), pp. 1–14. DOI: 10.1207/s15327809jls1301_1 (p. 33).

[BS14] M. Bada and A. Sasse. "Cyber Security Awareness Campaigns: Why Do They Fail to Change Behaviour?" In: (2014) (p. 9).

[BTP12] S. Bakkes, C. T. Tan, and Y. Pisan. "Personalised Gaming: A Motivation and Overview of Literature". In: *Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System*. Interactive Entertainment. IE ' 12. New York, USA: Association for Computing Machinery, 2012. DOI: 10.1145/2336727.2336731 (pp. 20, 26, 32).

[Bul04] S. Bull. "Supporting Learning with Open Learner Models". In: *Planning* 29.14 (2004), p. 1 (p. 21).

[Buna] Bundesamt für Sicherheit in der Informationstechnik. *Phishing E-Mails – Passwortdiebstahl durch Phishing*. Bundesamt für Sicherheit in der Informationstechnik. URL: https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/Cyber-Sicherheitslage/Methoden-der-Cyber-Kriminalitaet/Spam-Phishing-Co/Passwortdiebstahl-durch-Phishing/passwortdiebstahl-durch-phishing.html?nn=132228 (accessed on 17.06.2021) (pp. 2, 12).

[Bunb] Bundesamt für Sicherheit in der Informationstechnik. *Spam, Phishing & Co.* Spam, Phishing & Co. URL: https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/Cyber-Sicherheitslage/Methoden-der-Cyber-Kriminalitaet/Spam-Phishing-Co/spam-phishing-co.html;jsessionid=AE0F1F54CC8CFA5EED8976107E1FF2B6.internet471?nn=132348 (accessed on 17.06.2021) (pp. 2, 12).

[BV16] B. Bontchev and D. Vassileva. "Assessing Engagement in an Emotionally-Adaptive Applied Game". In: *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*. TEEM '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 747–754. DOI: 10.1145/3012430.3012602 (p. 27).

[Cab17] A. Caballero. “Chapter 33 - Security Education, Training, and Awareness”. In: *Computer and Information Security Handbook (Third Edition)*. Ed. by J. R. Vacca. Boston: Morgan Kaufmann, 2017, pp. 497–505. DOI: 10.1016/B978-0-12-803843-7.00033-8 (p. 9).

[Cam+07] R. Campbell, W. Robinson, J. Neelands, R. Hewston, and L. Mazzoli. “Personalised Learning: Ambiguities in Theory and Practice”. In: *British Journal of Educational Studies* 55.2 (2007), pp. 135–154. DOI: 10.1111/j.1467-8527.2007.00370.x (p. 29).

[Can+15] G. Canova, M. Volkamer, C. Bergmann, and B. Reinheimer. “NoPhish App Evaluation: Lab and Retention Study”. In: *NDSS Workshop on Usable Security 2015*. Network and Distributed System Security (NDSS) Symposium. USEC ’15. San Diego, California: Internet Society, 2015, pp. 1–10 (p. 134).

[CEW15] A. L. Compte, D. Elizondo, and T. Watson. “A renewed approach to serious games for cyber security”. In: *2015 7th International Conference on Cyber Conflict: Architectures in Cyberspace*. International Conference on Cyber Conflict: Architectures in Cyberspace. Tallinn, Estonia: IEEE, 2015, pp. 203–216. DOI: 10.1109/CYCON.2015.7158478 (p. 39).

[Chu13] E. F. Churchill. “Putting the Person Back into Personalization”. In: *Interactions* 20.5 (2013), pp. 12–15. DOI: 10.1145/2504847 (pp. 19, 21, 22).

[Cia06] R. B. Cialdini. *Influence: The Psychology of Persuasion, Revised Edition*. New York: New York: William Morrow, 2006 (pp. 1, 12, 145).

[CJ+18] G. CJ, S. Pandit, S. Vaddepalli, H. Tupsamudre, V. Banahatti, and S. Lodha. “PHISHY - A Serious Game to Train Enterprise Users on Phishing Awareness”. In: *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. Annual Symposium on Computer-Human Interaction in Play (Melbourne, VIC, Australia). CHI PLAY ’18 Extended Abstracts. New York, NY, USA: Association for Computing Machinery, 2018, pp. 169–181. DOI: 10.1145/3270316.3273042 (pp. 169, 170).

[CKV08] M. Cova, C. Kruegel, and G. Vigna. “There Is No Free Phish: An Analysis of "Free" and Live Phishing Kits”. In: *Proceedings of the 2nd Conference on USENIX Workshop on Offensive Technologies*. WOOT’08. San Jose, California, USA: USENIX Association, 2008, pp. 1–8 (p. 89).

[CMB11] S. Chiasson, M. Modi, and R. Biddle. “Auction Hero: The Design of a Game to Learn and Teach about Computer Security”. In: *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2011*. Ed. by C. Ho and M.-F. G. Lin. Honolulu, Hawaii, USA: Association for the Advancement of Computing in Education (AACE), 2011, pp. 2201–2206. URL: https://www.learntechlib.org/p/39053 (pp. 169, 170).

[CNS15] CNSS. *Committee on National Security Systems (CNSS) Glossary*. Glossary 2009. Committee on National Security Systems, 2015, p. 165. URL: https://rmf.org/wp-content/uploads/2017/10/CNSSI-4009.pdf (accessed on 13.08.2021) (p. 7).

[Con+07] B. D. Cone, C. E. Irvine, M. F. Thompson, and T. D. Nguyen. "A video game for cyber security training and awareness". In: *Computers & Security* 26.1 (2007), pp. 63–72. DOI: 10.1016/j.cose.2006.10.005 (pp. 169, 170).

[Con+12] 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". In: *Computers & Education* 59.2 (2012), pp. 661–686. DOI: 10.1016/j.compedu.2012.03.004 (p. 15).

[Csi97] M. Csikszentmihalyi. *Finding Flow: The Psychology of Engagement with Everyday Life*. Finding Flow: The Psychology of Engagement with Everyday Life. New York, NY, US: Basic Books, 1997, pp. ix, 181. ix, 181 (p. 25).

[DAJ11] D. Djaouti, J. Alvarez, and J.-P. Jessel. "Classifying Serious Games: The G/P/S Model". In: *Handbook of Research on Improving Learning and Motivation through Educational Games: Multidisciplinary Approaches*. IGI Global, 2011, pp. 118–136 (pp. 40, 41, 45–47, 129).

[Das+20] A. Das, S. Baki, A. El Aassal, R. Verma, and A. Dunbar. "SoK: A Comprehensive Reexamination of Phishing Research From the Security Perspective". In: *IEEE Communications Surveys Tutorials* 22.1 (2020), pp. 671–708. DOI: 10.1109/COMST.2019.2957750 (pp. 1, 12).

[DDC18] 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*. Ed. by S. Meiklejohn and K. Sako. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2018, pp. 160–179. DOI: 10.1007/978-3-662-58387-6_9 (pp. 1, 12).

[Det+11] S. Deterding, D. Dixon, R. Khaled, and L. Nacke. "From Game Design Elements to Gamefulness: Defining "Gamification"". In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. International Academic MindTrek Conference. MindTrek '11. Tampere, Finland: Association for Computing Machinery, 2011, pp. 9–15. DOI: 0.1145/2181037.2181040 (p. 14).

[DK16] R. Dey and J. Konert. "Content Generation for Serious Games". In: *Entertainment Computing and Serious Games: International GI-Dagstuhl Seminar 15283, Dagstuhl Castle, Germany, July 5-10, 2015, Revised Selected Papers*. Ed. by R. Dörner, S. Göbel, M. Kickmeier-Rust, M. Masuch, and K. Zweig. Cham: Springer International Publishing, 2016, pp. 174–188. DOI: 10.1007/978-3-319-46152-6_8 (pp. 28, 29, 32).

[DM19] V. Drury and U. Meyer. "Certified Phishing: Taking a Look at Public Key Certificates of Phishing Websites". In: *Proceedings of the Fifteenth Symposium on Usable Privacy and Security*. Symposium on Usable Privacy and Security. SOUPS' 19. Santa Clara, CA, USA: USENIX Association, 2019, pp. 211–223 (pp. 65, 105).

[Dör+16] R. Dörner, S. Göbel, W. Effelsberg, and J. Wiemeyer. "Introduction". In: *Serious Games: Foundations, Concepts and Practice*. Ed. by R. Dörner, S. Göbel, W. Effelsberg, and J. Wiemeyer. Cham: Springer International Publishing, 2016, pp. 1–34. DOI: 10.1007/978-3-319-40612-1_1 (pp. 14, 15, 17).

[DTH06] R. Dhamija, J. D. Tygar, and M. Hearst. “Why Phishing Works”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Conference on Human Factors Computing Systems. CHI ’06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 581–590. DOI: 10.1145/1124772.1124861 (pp. 1, 12).

[Eck12] C. Eckert. *IT-Sicherheit: Konzepte - Verfahren - Protokolle*. 7., überarb. und erw. Aufl. München: Oldenbourg-Verl, 2012. 1004 pp. (p. 7).

[ERS18] M. Ehlenz, R. Roepke, and U. Schroeder. “Towards Multi-touch Learning Applications in Collaborative Education”. In: *PerDis ’18: Proceedings of the 7th ACM International Symposium on Pervasive Displays*. PerDis ’18: The International Symposium on Pervasive Displays. New York: Association for Computing Machinery, 2018, pp. 1–2. DOI: 10.1145/3205873.3210709 (pp. 18, 36).

[ES18] Y. Elsayed and A. Shosha. “Large Scale Detection of IDN Domain Name Masquerading”. In: *2018 APWG Symposium on Electronic Crime Research (eCrime)*. APWG Symposium on Electronic Crime Research. eCrime ’18. New York, NY, USA: IEEE, 2018, pp. 1–11. DOI: 10.1109/ECRIME.2018.8376212 (p. 56).

[FA19] M. Fernando and N. A. G. Arachchilage. “Why Johnny Can’t Rely on Anti-Phishing Educational Interventions to Protect Himself against Contemporary Phishing Attacks?” In: *ACIS 2019 Proceedings*. Australasian Conference on Information Systems. Vol. abs/2004.13262. ACIS ’19. Perth, Australia, 2019, pp. 395–405. URL: https://arxiv.org/abs/2004.13262 (p. 12).

[Fed] Federal Bureau of Investigation. *Spoofing and Phishing*. Federal Bureau of Investigation. URL: https://www.fbi.gov/scams-and-safety/common-scams-and-crimes/spoofing-and-phishing (accessed on 17.06.2021) (pp. 2, 12).

[FH03] A. P. Field and G. Hole. *How to Design and Report Experiments*. London, UK: Sage, 2003. 384 pp. (p. 36).

[FM14] S. Furnell and L. Moore. “Security Literacy: The Missing Link in Today’s Online Society?” In: *Computer Fraud & Security* 2014.5 (2014), pp. 12–18. DOI: 10.1016/S1361-3723(14)70491-9 (p. 8).

[FM92] N. D. Fleming and C. Mills. “Not Another Inventory, Rather a Catalyst for Reflection”. In: *To Improve the Academy* 11.1 (1992), pp. 137–155. DOI: 10.1002/j.2334-4822.1992.tb00213.x (pp. 24, 25).

[FMS19] D. Filipczuk, C. Mason, and S. Snow. “Using a Game to Explore Notions of Responsibility for Cyber Security in Organisations”. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. Conference on Human Factors in Computing Systems. CHI EA ’19. New York, NY, USA: Association for Computing Machinery, 2019. DOI: 10.1145/3290607.3312846 (pp. 169, 170).

[FP06] H. Fan and M. S. Poole. “What Is Personalization? Perspectives on the Design and Implementation of Personalization in Information Systems”. In: *Journal of Organizational Computing and Electronic Commerce* 16.3-4 (2006), pp. 179–202. DOI: 10.1080/10919392.2006.9681199 (pp. 18, 19, 22).

[Fre+16] M. Freire, Á. Serrano-Laguna, B. M. Iglesias, I. Martínez-Ortiz, P. Moreno-Ger, and B. Fernández-Manjón. "Game Learning Analytics: Learning Analytics for Serious Games". In: *Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy*. Ed. by M. J. Spector, B. B. Lockee, and M. D. Childress. Cham: Springer International Publishing, 2016, pp. 1–29. DOI: 10.1007/978-3-319-17727-4_21-1 (pp. 18, 38).

[Fre+19] S. Frey, A. Rashid, P. Anthonysamy, M. Pinto-Albuquerque, and S. A. Naqvi. "The Good, the Bad and the Ugly: A Study of Security Decisions in a Cyber-Physical Systems Game". In: *IEEE Transactions on Software Engineering* 45.5 (2019), pp. 521–536. DOI: 10.1109/TSE.2017.2782813 (pp. 169, 170).

[Fru20] J. Fruhlinger. *The CIA Triad: Definition, Components and Examples*. CSO Online. 2020. URL: https://www.csoonline.com/article/3519908/the-cia-triad-definition-components-and-examples.html (accessed on 08.09.2020) (p. 7).

[FS88] R. M. Felder and L. K. Silverman. "Learning and Teaching Styles in Engineering Education". In: *Engineering Education* 78.7 (1988), pp. 674–681 (p. 24).

[Fun+08] C. C. Fung, V. Khera, A. Depickere, P. Tantatsanawong, and P. Boonbrahm. "Raising Information Security Awareness in Digital Ecosystem with Games - a Pilot Study in Thailand". In: *2008 2nd IEEE International Conference on Digital Ecosystems and Technologies*. International Conference on Digital Ecosystems and Technologies. DEST '08. New York, NY, USA: IEEE, 2008, pp. 375–380. DOI: 10.1109/DEST.2008.4635145 (p. 49).

[Gar+07] S. Garera, N. Provos, M. Chew, and A. D. Rubin. "A Framework for Detection and Measurement of Phishing Attacks". In: *Proceedings of the 2007 ACM Workshop on Recurring Malcode*. Workshop on Recurring Malcode. WORM '07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 1–8. DOI: 10.1145/1314389.1314391 (p. 12).

[GB09] M. J. Grant and A. Booth. "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies: A Typology of Reviews, *Maria J. Grant & Andrew Booth*". In: *Health Information & Libraries Journal* 26.2 (2009), pp. 91–108. DOI: 10.1111/j.1471-1842.2009.00848.x (pp. 35, 133).

[Gee07] J. P. Gee. *What Video Games Have to Teach Us about Learning and Literacy. Second Edition: Revised and Updated Edition*. Palgrave Macmillan, 2007 (pp. 13, 15).

[Gey19] J. Geywitz. ""What the Hack?" - Konzeption und Implementierung eines erweiterbaren und adaptiven Serious Game zur Verbesserung von Information Security Awareness". MA thesis. Düsseldorf: Hochschule Düsseldorf, University of Applied Sciences, 2019. 79 pp. (pp. 58, 169, 170).

[GKG15] F. Giannakas, G. Kambourakis, and S. Gritzalis. "CyberAware: A mobile game-based app for cybersecurity education and awareness". In: *2015 International Conference on Interactive Mobile Communication Technologies and Learning*. International Conference on Interactive Mobile Communication Technologies and Learning. IMCL '15. Thessaloniki, Greece:

IEEE, 2015, pp. 54–58. DOI: 10.1109/IMCTL.2015.7359553 (pp. 58, 169, 170).

[Göb+10] S. Göbel, S. Hardy, V. Wendel, F. Mehm, and R. Steinmetz. "Serious Games for Health: Personalized Exergames". In: *Proceedings of the 18th ACM International Conference on Multimedia*. ACM International Conference on Multimedia. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 1663–1666. DOI: 10.1145/1873951.1874316 (pp. 28, 32).

[GP13] M. Gondree and Z. N. J. Peterson. "Valuing Security by Getting [d0x3d!]: Experiences with a Network Security Board Game". In: *6th Workshop on Cyber Security Experimentation and Test*. Workshop on Cyber Security Experimentation and Test. CSET '13. Washington, D.C.: USENIX Association, 2013 (pp. 169, 170).

[GW16] S. Göbel and V. Wendel. "Personalization and Adaptation". In: *Serious Games: Foundations, Concepts and Practice*. Ed. by R. Dörner, S. Göbel, W. Effelsberg, and J. Wiemeyer. Cham: Springer International Publishing, 2016, pp. 161–210. DOI: 10.1007/978-3-319-40612-1_7 (pp. 26, 29, 31).

[GWD17] S. Goel, K. Williams, and E. Dincelli. "Got Phished? Internet Security and Human Vulnerability". In: *Journal of the Association for Information Systems* 18.1 (2017). DOI: 10.17705/1jais.00447 (p. 12).

[HAB16] M. Hendrix, A. Al-Sherbaz, and V. Bloom. "Game Based Cyber Security Training: Are Serious Games Suitable for Cyber Security Training?" In: *International Journal of Serious Games* 3.1 (2016), pp. 53–61. DOI: 10.17083/ijsg.v3i1.107 (pp. 39, 40).

[Heb+17] A. J. Hebert, C. O. Reynolds, K. J. Stack, and R. C. Lindsay. *Lock_Out: A Cybersecurity MQP and Game*. Final Report. Worcester, MA, USA: Worcester Polytechnic Institute, 2017, p. 32. URL: https://digitalcommons.wpi.edu/mqp-all/1871/ (accessed on 30.03.2020) (pp. 169, 170).

[Her+07] J. Herrington, S. McKenney, T. Reeves, and R. Oliver. "Design-Based Research and Doctoral Students: Guidelines for Preparing a Dissertation Proposal". In: EdMedia + Innovate Learning. Association for the Advancement of Computing in Education (AACE), 2007, pp. 4089–4097. URL: https://www.learntechlib.org/primary/p/25967/ (accessed on 20.12.2021) (p. 33).

[HGG15] M. L. Hale, R. F. Gamble, and P. Gamble. "CyberPhishing: A Game-Based Platform for Phishing Awareness Testing". In: *2015 48th Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. Vol. 48. Kauai, HI, USA: IEEE, 2015, pp. 5260–5269. DOI: 10.1109/HICSS.2015.670 (pp. 169, 170).

[Hig05] S. D. Hight. "The Importance of a Security, Education, Training and Awareness Program (November 2005)". In: 2005 (pp. 8, 9).

[Hoc19] N. Hocine. "Personalized Serious Games for Self-regulated Attention Training". In: *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. Conference on User Modeling, Adaptation and Personalization. UMAP'19 Adjunct. New York, NY,

USA: Association for Computing Machinery, 2019, pp. 251–255. DOI: 10.1145/3314183.3323458 (p. 28).

[HRH20] B. D. Homer, C. Raffaele, and H. Henderson. "Games as Playful Learning: Implications of Developmental Theory for Game-Based Learning". In: *Handbook of Game-Based Learning*. Cambridge, MA, US: The MIT Press, 2020, pp. 25–52 (p. 15).

[Huy+17] D. Huynh, P. Luong, H. Iida, and R. Beuran. "Design and Evaluation of a Cybersecurity Awareness Training Game". In: *Entertainment Computing – ICEC 2017*. International Conference on Entertainment Computing. Ed. by N. Munekata, I. Kunita, and J. Hoshino. Cham: Springer International Publishing, 2017, pp. 183–188. DOI: 10.1007/978-3-319-66715-7_19. URL: https://link.springer.com/chapter/10.1007/978-3-319-66715-7_19 (pp. 58, 169, 170).

[HW18] H. Hu and G. Wang. "End-to-End Measurements of Email Spoofing Attacks". In: *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, 2018, pp. 1095–1112 (p. 56).

[Hwa+12] G.-J. Hwang, H.-Y. Sung, C.-M. Hung, I. Huang, and C.-C. Tsai. "Development of a Personalized Educational Computer Game Based on Students' Learning Styles". In: *Educational Technology Research and Development* 60.4 (2012), pp. 623–638. DOI: 10.1007/s11423-012-9241-x (p. 27).

[IB18] H. Ismail and B. Belkhouche. "A Reusable Software Architecture for Personalized Learning Systems". In: *2018 International Conference on Innovations in Information Technology*. International Conference on Innovations in Information Technology. IIT '18. New York, NY, USA: IEEE, 2018, pp. 105–110. DOI: 10.1109/INNOVATIONS.2018.8605997 (pp. 20, 21, 134).

[ISO18] ISO/IEC. *ISO/IEC 27000:2018(En), Information Technology — Security Techniques — Information Security Management Systems — Overview and Vocabulary*. 2018. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:27000:ed-5:v1:en (accessed on 08.09.2020) (p. 7).

[ITA05] C. E. Irvine, M. F. Thompson, and K. Allen. "CyberCIEGE: Gaming for Information Assurance". In: *IEEE Security Privacy* 3.3 (2005), pp. 61–64. DOI: 10.1109/MSP.2005.64 (pp. 49, 61).

[Jär08] A. Järvinen. *Games without Frontiers: Theories and Methods for Game Studies and Design*. Tampere University Press, 2008. URL: https://urn.fi/urn:isbn:978-951-44-7252-7 (accessed on 02.03.2022) (p. 14).

[JI16] D. Jacobson and J. Idziorek. *Computer Security Literacy: Staying Safe in a Digital World*. CRC Press, 2016 (p. 8).

[Jon+10] J. Jones, X. Yuan, E. Carr, and H. Yu. "A Comparative Study of CyberCIEGE Game and Department of Defense Information Assurance Awareness Video". In: *Proceedings of the IEEE SoutheastCon 2010*. IEEE SoutheastCon. SoutheastCon '10. New York, NY, USA: IEEE, 2010, pp. 176–180. DOI: 10.1109/SECON.2010.5453895 (p. 49).

[KA08] M. D. Kickmeier-Rust and D. Albert. “The ELEKTRA Ontology Model: A Learner-Centered Approach to Resource Description”. In: *Advances in Web Based Learning*. International Conference on Web-Based Learning. ICWL ’07. Berlin, Heidelberg: Springer, 2008, pp. 78–89. DOI: 10.1007/978-3-540-78139-4_8 (p. 27).

[KA10] M. D. Kickmeier-Rust and D. Albert. “Micro-Adaptivity: Protecting Immersion in Didactically Adaptive Digital Educational Games”. In: *Journal of Computer Assisted Learning* 26.2 (2010), pp. 95–105. DOI: 10.1111/j.1365-2729.2009.00332.x (p. 121).

[Kar+17] S. Karwatzki, O. Dytynko, M. Trenz, and D. Veit. “Beyond the Personalization–Privacy Paradox: Privacy Valuation, Transparency Features, and Service Personalization”. In: *Journal of Management Information Systems* 34.2 (2017), pp. 369–400. DOI: 10.1080/07421222.2017.1334467 (p. 19).

[Kas21] Kaspersky. *Spam and Phishing in Q3 2021*. 2021. URL: https://securelist.com/spam-and-phishing-in-q3-2021/ (accessed on 10.11.2021) (p. 1).

[Kat+17] E. Katsadouros, D. Kogias, L. Toumanidis, C. Chatzigeorgiou, and C. Z. Patrikakis. “Teaching network security through a scavenger hunt game”. In: *2017 IEEE Global Engineering Education Conference*. Global Engineering Education Conference. EDUCON ’17. Athens, Greece: IEEE, 2017, pp. 1802–1805. DOI: 10.1109/EDUCON.2017.7943094 (pp. 169, 170).

[KGA08] M. D. Kickmeier-Rust, S. Göbel, and D. Albert. “80Days: Melding Adaptive Educational Technology and Adaptive and Interactive Storytelling in Digital Educational Games”. In: *Proceedings of the First International Workshop on Story-Telling and Educational Games*. International Workshop on Story-Telling and Educational Games. Vol. 386. STEG’ 08. Maastricht, Netherlands: CEUR, 2008, pp. 11–18 (pp. 27, 31).

[Kic+06] M. Kickmeier-Rust, D. Schwarz, D. Albert, D. Verpoorten, J.-L. Castaigne, and M. Bopp. “The ELEKTRA Project: Towards a New Learning Experience”. In: *Proceedings of the 2nd Symposium of the WG HCI&UE of the Austrian Computer Society*. Symposium of the WG HCI &UE of the Austrian Computer Society. Vienna, Austria: Österreichische Computer Gesellschaft, 2006, pp. 19–48. DOI: 10.13140/2.1.2272.8646 (pp. 27, 31).

[Kic+07a] M. Kickmeier-Rust, D. Albert, C. Hockemeyer, and T. Augustin. “Not Breaking the Narrative: Individualized Competence Assessment in Educational Games”. In: *Proceedings of the European Conference on Games Based Learning*. European Conference on Games Based Learning. Ed. by D. Remenyi. ECGBL ’07. Paisley, UK: Curran, 2007, pp. 161–168 (p. 27).

[Kic+07b] M. D. Kickmeier-Rust, N. Peirce, O. Conlan, D. Schwarz, D. Verpoorten, and D. Albert. “Immersive Digital Games: The Interfaces for Next-Generation E-Learning?” In: *Universal Access in Human-Computer Interaction. Applications and Services*. International Conference on Universal Access in Human-Computer Interaction. Ed. by C. Stephanidis. UAHCI ’07. Berlin, Heidelberg: Springer, 2007, pp. 647–656. DOI: 10.1007/978-3-540-73283-9_71 (p. 27).

[Kic+08] M. D. Kickmeier-Rust, C. Hockemeyer, D. Albert, and T. Augustin. "Micro Adaptive, Non-invasive Knowledge Assessment in Educational Games". In: *2008 Second IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning*. International Conference on Digital Game and Intelligent Toy Enhanced Learning. DIGITEL '08. Banff, Alberta, Canada: IEEE, 2008, pp. 135–137. DOI: 10.1109/DIGITEL.2008.10 (p. 27).

[Kic+11] M. D. Kickmeier-Rust, E. Mattheiss, C. Steiner, and D. Albert. "A Psycho-Pedagogical Framework for Multi-Adaptive Educational Games". In: *International Journal of Game-Based Learning (IJGBL)* 1.1 (2011), pp. 45–58. DOI: 10.4018/ijgbl.2011010104 (p. 27).

[KIJ13] M. Khonji, Y. Iraqi, and A. Jones. "Phishing Detection: A Literature Survey". In: *IEEE Communications Surveys Tutorials* 15.4 (2013), pp. 2091–2121. DOI: 10.1109/SURV.2013.032213.00009 (p. 10).

[Kin+17] P. Kintis et al. "Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse". In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. Conference on Computer and Communications Security. CCS '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 569–586. DOI: 10.1145/3133956.3134002 (pp. 13, 65, 82, 88).

[Kir17] P. A. Kirschner. "Stop Propagating the Learning Styles Myth". In: *Computers & Education* 106 (2017), pp. 166–171. DOI: 10.1016/j.compedu.2016.12.006 (p. 24).

[KKS11] S. Kelle, R. Klemke, and M. Specht. "Design Patterns for Learning Games". In: *International Journal of Technology Enhanced Learning* 3.6 (2011), pp. 555–569. DOI: 10.1504/IJTEL.2011.045452 (p. 17).

[Kot04] C. R. Kothari. *Research Methodology: Methods and Techniques*. New Age International, 2004. 30 pp. (p. 34).

[Kra02] D. R. Krathwohl. "A Revision of Bloom's Taxonomy: An Overview". In: *Theory Into Practice* 41.4 (2002), pp. 212–218. DOI: 10.1207/s15430421tip4104_2 (pp. 35, 36, 50, 53, 66).

[KS12] I. Kirlappos and M. A. Sasse. "Security Education against Phishing: A Modest Proposal for a Major Rethink". In: *IEEE Security Privacy* 10.2 (2012), pp. 24–32. DOI: 10.1109/MSP.2011.179 (pp. 2, 12).

[KSvK21] J. Krath, L. Schürmann, and H. F. O. von Korflesch. "Revealing the Theoretical Basis of Gamification: A Systematic Review and Analysis of Theory in Research on Gamification, Serious Games and Game-Based Learning". In: *Computers in Human Behavior* 125.106963 (2021). DOI: 10.1016/j.chb.2021.106963 (p. 16).

[Küh18] D. Kühn. "Ein Modul zur Generierung von Feedback und Hinweisen in Multitouch-Lernspielen". BA thesis. Aachen: RWTH Aachen University, 2018. 64 pp. URL: https://publications.rwth-aachen.de/record/728905 (p. 18).

[Kul19] V. K. Kulkarni. "Basic Cybersecurity Awareness Through Gaming". MA thesis. Fargo: North Dakota State University, 2019. 51 pp. URL: https://library.ndsu.edu/ir/handle/10365/29222 (accessed on 30.03.2020) (pp. 169, 170).

[Kum+08] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. "Lessons from a Real World Evaluation of Anti-Phishing Training". In: *2008 eCrime Researchers Summit*. eCrime Researchers Summit. eCrime '08. Atlanta, Georgia, USA: IEEE, 2008, pp. 1–12. DOI: 10.1109/ECRIME.2008.4696970 (p. 121).

[KW18] J. A. König and M. R. Wolf. "GHOST: An Evaluated Competence Developing Game for Cybersecurity Awareness Training". In: *International Journal on Advances in Security* 11 (3 & 4 2018), pp. 274–287 (pp. 58, 169, 170).

[Lan14] R. N. Landers. "Developing a Theory of Gamified Learning: Linking Serious Games and Gamification of Learning". In: *Simulation & Gaming* 45.6 (2014), pp. 752–768. DOI: 10.1177/1046878114563660 (pp. 16, 17).

[Las14] E. E. Lastdrager. "Achieving a Consensual Definition of Phishing Based on a Systematic Review of the Literature". In: *Crime Science* 3.1 (2014), p. 10. DOI: 10.1186/s40163-014-0009-y (pp. 1, 10).

[Lin+19] T. Lin et al. "Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content". In: *ACM Transactions on Computer-Human Interaction* 26.5 (2019), 32:1–32:28. DOI: 10.1145/3336141 (p. 121).

[Lop+18] I. Lopes, Y. Morenets, P. R. M. Inácio, and F. Silva. "Cyber-Detective: A Game for Cyber Crime Prevention". In: *Proceedings of Play2Learn*. Play2Learn. Lisbon, Portugal, 2018, pp. 175–191. URL: http://gamilearning.ulusofona.pt/play2learn-2018-proceedings/ (accessed on 30.03.2020) (pp. 169, 170).

[LR08] E. L.-C. Law and M. Rust-Kickmeier. "80Days: Immersive Digital Educational Games with Adaptive Storytelling". In: *Proceedings of the First International Workshop on Story-Telling and Educational Games (STEG'08) The Power of Narration and Imagination in Technology Enhanced Learning*. Story-Telling and Educational Games. Vol. 386. STEG '08. Maastricht, Netherlands: CEUR, 2008, pp. 56–62 (p. 121).

[Lu18] Y. Lu. "CyberCraft, a security serious game". MA thesis. Torino, Italy: Politecnico di Torino, 2018. 92 pp. URL: https://webthesis.biblio.polito.it/9474/ (accessed on 30.03.2020) (pp. 58, 169, 170).

[LW19] K. C. Li and B. T.-M. Wong. "How Learning Has Been Personalised: A Review of Literature from 2009 to 2018". In: *Blended Learning: Educational Innovation for Personalized Learning*. Ed. by S. K. S. Cheung, L.-K. Lee, I. Simonova, T. Kozel, and L.-F. Kwok. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 72–81. DOI: 10.1007/978-3-030-21562-0_6 (pp. 29, 30).

[LW21] K. C. Li and B. T.-M. Wong. "Features and Trends of Personalised Learning: A Review of Journal Publications from 2001 to 2018". In: *Interactive Learning Environments* 29.2 (2021), pp. 182–195. DOI: 10.1080/10494820.2020.1811735 (pp. 29, 30).

[MAB17] G. Misra, N. A. G. Arachchilage, and S. Berkovsky. "Phish Phinder: A Game Design Approach to Enhance User Confidence in Mitigating Phishing Attacks". In: *Proceedings of the Eleventh International Symposium on Human Aspects of Information Security & Assurance*. International Symposium on Human Aspects of Information Security & Assurance. Ed. by

S. Furnell and N. L. Clarke. HAISA '17. Adelaide, Australia, 2017, pp. 41–51 (pp. 169, 170).

[May14] R. E. Mayer. *Computer Games for Learning: An Evidence-Based Approach*. Computer Games for Learning: An Evidence-Based Approach. Cambridge, MA, US: MIT Press, 2014. 281 pp. (p. 13).

[MC05] D. R. Michael and S. L. Chen. *Serious Games: Games That Educate, Train, and Inform*. Muska & Lipman/Premier-Trade, 2005 (p. 14).

[McG11] J. McGonigal. *Reality Is Broken: Why Games Make Us Better and How They Can Change the World*. Penguin, 2011 (p. 13).

[MDZ14] V. Montani, M. De Filippo De Grazia, and M. Zorzi. "A New Adaptive Videogame for Training Attention and Executive Functions: Design Principles and Initial Validation". In: *Frontiers in Psychology* 5 (2014), p. 409. DOI: 10.3389/fpsyg.2014.00409 (p. 28).

[Mey06] S. Meyers. "Introduction to Phishing". In: *Phishing and Countermeasures*. John Wiley & Sons, Ltd, 2006, pp. 1–29. DOI: 10.1002/9780470086100.ch1 (p. 10).

[MG08] D. K. McGrath and M. Gupta. "Behind Phishing: An Examination of Phisher Modi Operandi". In: *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*. 2008, pp. 1–8 (p. 56).

[MJ10] R. E. Mayer and C. I. Johnson. "Adding Instructional Features That Promote Learning in a Game-Like Environment". In: *Journal of Educational Computing Research* 42.3 (2010), pp. 241–265. DOI: 10.2190/EC.42.3.a (p. 13).

[ML03] J. Meyer and R. Land. "Threshold Concepts and Troublesome Knowledge: Linkages to Ways of Thinking and Practising within the Disciplines". In: *Improving Student Learning Theory and Practice - 10 Years on: Proceedings of the 2002 10th International Symposium Improving Student Learning*. International Symposium on Improving Student Learning. Ed. by C. Rust. Vol. 10. Oxford, UK: Oxford Centre for Staff & Learning Development, 2003, pp. 412–424 (p. 24).

[MPJ18] J. Mikka-Muntuumo, A. Peters, and H. Jazri. "CyberBullet - Share Your Story: An Interactive Game for Stimulating Awareness on the Harm and Negative Effects of the Internet". In: *Proceedings of the Second African Conference for Human Computer Interaction: Thriving Communities*. African Conference for Human Computer Interaction. New York, NY, USA: Association for Computing Machinery, 2018, pp. 287–290. DOI: 10.1145/3283458.3283482 (pp. 169, 170).

[MvNvS10] T. Monk, J. van Niekerk, and R. von Solms. "Sweetening the Medicine: Educating Users about Information Security by Means of Game Play". In: *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*. Annual Research Conference of the South African Institute of Computer Scientists and Information. SAICSIT '10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 193–200. DOI: 10.1145/1899503.1899525 (pp. 169, 170).

[N26] N26 GmbH. *Preventing Phishing Attacks on Your Bank Account*. Preventing phishing attacks on your bank account. URL: https://n26.com/en-eu/blog/preventing-phishing-attacks-on-your-bank-account (accessed on 17.06.2021) (pp. 2, 12).

[NR09] J. Niehaus and M. O. Riedl. "Scenario Adaptation: An Approach to Customizing Computer-Based Training Games and Simulations". In: *Proceedings of the AIED 2009 Workshop on Intelligent Educational Games*. International Conference on Artificial Intelligence in Education. Vol. 3. AIED '09. Brighton, UK, 2009, pp. 89–98 (p. 26).

[Oes+18] 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: *2018 APWG Symposium on Electronic Crime Research*. Symposium on Electronic Crime Research. eCrime '18. New York, 2018, pp. 1–12. DOI: 10.1109/ECRIME.2018.8376206 (pp. 11, 65, 82, 88).

[Oes+19] 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: *2019 IEEE Symposium on Security and Privacy (SP)*. Symposium on Security and Privacy. SP' 19. New York: IEEE, 2019, pp. 1344–1361. DOI: 10.1109/SP.2019.00049 (pp. 1, 11).

[Oes+20] A. Oest et al. "Sunrise to Sunset: Analyzing the End-to-end Life Cycle and Effectiveness of Phishing Attacks at Scale". In: *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Security Symposium. USENIX Security '20. USENIX Association, 2020, pp. 361–377. URL: https://www.usenix.org/conference/usenixsecurity20/presentation/oest-sunrise (accessed on 10.06.2021) (pp. 1, 11).

[Ola+14] M. Olano et al. "SecurityEmpire: Development and Evaluation of a Digital Game to Promote Cybersecurity Education". In: *Proceedings of 2014 USENIX Summit on Gaming, Games, and Gamification in Security Education*. USENIX Summit on Gaming, Games, and Gamification in Security Education. 3GSE '14. San Diego, CA, USA: USENIX Association, 2014 (pp. 169, 170).

[Oli+17] D. Oliveira et al. "Dissecting Spear Phishing Emails for Older vs Young Adults: On the Interplay of Weapons of Influence and Life Domains in Predicting Susceptibility to Phishing". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Conference on Human Factors in Computing Systems. CHI '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 6412–6424. DOI: 10.1145/3025453.3025831 (p. 105).

[OOT17] R. Orji, K. Oyibo, and G. F. Tondello. "A Comparison of System-Controlled and User-Controlled Personalization Approaches". In: *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*. Conference on User Modeling, Adaptation and Personalization. UMAP '17. Bratislava, Slovakia: Association for Computing Machinery, 2017, pp. 413–418. DOI: 10.1145/3099023.3099116 (p. 20).

[Orj+13] R. Orji, R. L. Mandryk, J. Vassileva, and K. M. Gerling. "Tailoring Persuasive Health Games to Gamer Type". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Conference on Human Factors in Computing Systems. CHI '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 2467–2476. DOI: 10.1145/2470654.2481341 (p. 28).

[OZ15] A.-S. T. Olanrewaju and N. H. Zakaria. "Social engineering awareness game (SEAG): an empirical evaluation of using game towards improving information security awareness". In: *Proceedings of the 5th International Conference on Computing and Informatics*. International Conference on Computing and Informatics. ICOCI '15. Istanbul, Turkey, 2015, pp. 187–193 (pp. 169, 170).

[Par+15] G. Paré, M.-C. Trudel, M. Jaana, and S. Kitsiou. "Synthesizing Information Systems Knowledge: A Typology of Literature Reviews". In: *Information & Management* 52.2 (2015), pp. 183–199. DOI: 10.1016/j.im.2014.08.008 (pp. 35, 133).

[Pat+18] M. R. Pattinson et al. "Adapting Cyber-Security Training to Your Employees". In: *Twelfth International Symposium on Human Aspects of Information Security & Assurance, HAISA 2018, Dundee, Scotland, UK, August 29-31, 2018, Proceedings*. International Symposium on Human Aspects of Information Security & Assurance. Ed. by N. L. Clarke and S. Furnell. HAISA '18. Dundee, UK: University of Plymouth, 2018, pp. 67–79 (pp. 25, 26, 31).

[Pay] PayPal. *Recognizing Suspicious Activity - PayPal*. Recognize fraudulent emails and websites. URL: https://www.paypal.com/us/webapps/mpp/security/suspicious-activity (accessed on 17.06.2021) (pp. 2, 12).

[PDC10] V. Pastor, G. Díaz, and M. Castro. "State-of-the-art simulation systems for information security education, training and awareness". In: *IEEE EDUCON 2010 Conference*. Global Education Engineering. EDUCON '10. Madrid: IEEE, 2010, pp. 1907–1916. DOI: 10.1109/EDUCON.2010.5492435 (p. 39).

[PHK15] J. L. Plass, B. D. Homer, and C. K. Kinzer. "Foundations of Game-Based Learning". In: *Educational Psychologist* 50.4 (2015), pp. 258–283. DOI: 10.1080/00461520.2015.1122533 (pp. 13, 16).

[Pla+19] J. L. Plass, B. D. Homer, S. Pawar, C. Brenner, and A. P. MacNamara. "The Effect of Adaptive Difficulty Adjustment on the Effectiveness of a Game to Develop Executive Function Skills for Learners of Different Ages". In: *Cognitive Development* 49 (2019), pp. 56–67. DOI: 10.1016/j.cogdev.2018.11.006 (p. 27).

[Pla+20] J. L. Plass, B. D. Homer, R. E. Mayer, and C. K. Kinzer. "Theoretical Foundations of Game-Based and Playful Learning." In: *Handbook of Game-Based Learning*. Cambridge, MA, US: The MIT Press, 2020, pp. 3–24 (p. 14).

[Pre03] M. Prensky. "Digital Game-Based Learning". In: *Computers in Entertainment* 1.1 (2003), pp. 21–21. DOI: 10.1145/950566.950596 (pp. 13, 15).

[Qua21] F. Quayyum. “Cyber Security Education for Children Through Gamification: Challenges and Research Perspectives”. In: *Methodologies and Intelligent Systems for Technology Enhanced Learning, 10th International Conference. Workshops*. International Conference in Methodologies and Intelligent Systems for Techhnology Enhanced Learning. Ed. by Z. Kubincová, L. Lancia, E. Popescu, M. Nakayama, V. Scarano, and A. B. Gil. MIS4TEL ’20. Cham: Springer International Publishing, 2021, pp. 258–263. DOI: 10.1007/978-3-030-52287-2_26 (p. 23).

[Res08] P. Resnick. *Internet Message Format*. RFC 5322. 2008. DOI: 10.17487/RFC5322. URL: https://rfc-editor.org/rfc/rfc5322.txt (p. 56).

[Rey+20] J. Reynolds et al. “Measuring Identity Confusion with Uniform Resource Locators”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Conference on Human Factors in Computing Systems. CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–12. DOI: 10.1145/3313831.3376298 (pp. 1, 12, 13).

[RL16] A. Rieb and U. Lechner. “Operation Digital Chameleon: Towards an Open Cybersecurity Method”. In: *Proceedings of the 12th International Symposium on Open Collaboration*. International Symposium on Open Collaboration. OpenSym ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1–10. DOI: 10.1145/2957792.2957800 (pp. 169, 170).

[RLA14] R. Raman, A. Lal, and K. Achuthan. “Serious Games Based Approach to Cyber Security Concept Learning: Indian Context”. In: *2014 International Conference on Green Computing Communication and Electrical Engineering*. International Conference on Green Computing Communication and Electrical Engineering. ICGCCEE ’14. Coimbatore, India: IEEE, 2014, pp. 1–5. DOI: 10.1109/ICGCCEE.2014.6921392 (p. 49).

[Rob+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: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. Conference on Computer and Communications Security. CCS ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 2489–2504. DOI: 10.1145/3319535.3363188 (pp. 13, 65, 82, 88).

[Sch15] B. Schneier. *Secrets and Lies: Digital Security in a Networked World*. John Wiley & Sons, 2015 (pp. 1, 11).

[SDC13] M. Seif El-Nasr, A. Drachen, and A. Canossa. “Introduction”. In: *Game Analytics: Maximizing the Value of Player Data*. Ed. by M. Seif El-Nasr, A. Drachen, and A. Canossa. London: Springer, 2013, pp. 3–12. DOI: 10.1007/978-1-4471-4769-5_1 (p. 18).

[SEC21] SECUSO. *KIT - SECUSO Forschung - Ergebnisse - S&P Awareness/Education/Training - Sichere Kommunikation*. Security und Privacy (S&P) Awareness-/Education- und Trainingsmaßnahmen zur Ermöglichung einer sicheren Kommunikation. 2021. URL: https://secuso.aifb.kit.edu/99.php (accessed on 17.06.2021) (pp. 2, 12).

[SH17] T. Seitz and H. Hussmann. "PASDJO: Quantifying Password Strength Perceptions with an Online Game". In: *Proceedings of the 29th Australian Conference on Computer-Human Interaction*. Australian Conference on Computer-Human Interaction. OZCHI '17. New York, NY, USA: ACM, 2017, pp. 117–125. DOI: 10.1145/3152771.3152784 (p. 49).

[She+07] S. Sheng et al. "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* (Pittsburgh, Pennsylvania, USA). SOUPS '07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 88–99. DOI: 10.1145/1280680.1280692 (pp. 2, 57, 58, 77, 134, 169, 170).

[She+09] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang. "An Empirical Analysis of Phishing Blacklists". In: *Proceedings of the 6th Conference on Email and Anti-Spam*. Conference on Email and Anti-Spam. CEAS '09. Mountain View, CA, USA: Carnegie Mellon University, 2009. DOI: 10.1184/R1/6469805.v1 (pp. 1, 11).

[Sic08] M. Sicart. "Defining Game Mechanics". In: *Game Studies* 8.2 (2008). URL: http://gamestudies.org/0802/articles/sicart (accessed on 02.03.2022) (p. 14).

[Sie10] G. Siemens. *1st International Conference on Learning Analytics and Knowledge*. Learning Analytics & Knowledge: February 27-March 1, 2011 in Banff, Alberta. 2010. URL: https://tekri.athabascau.ca/analytics/ (accessed on 27.01.2022) (p. 18).

[SKA09] C. M. Steiner, M. D. Kickmeier-Rust, and D. Albert. "Little Big Difference: Gender Aspects and Gender-Based Adaptation in Educational Games". In: *Learning by Playing. Game-based Education System Design and Development*. Ed. by M. Chang, R. Kuo, Kinshuk, G.-D. Chen, and M. Hirose. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 150–161. DOI: 10.1007/978-3-642-03364-3_20 (p. 27).

[SLJ17] A. Samuels, F. Li, and C. Justice. "Applying Rating Systems to Challenge Based Cybersecurity Education". In: *2017 International Conference on Computing, Networking and Communications (ICNC)*. International Conference on Computing, Networking and Communications. ICNC '17. Silicon Valley, CA, USA: IEEE, 2017, pp. 819–824. DOI: 10.1109/ICCNC.2017.7876237 (pp. 25, 26, 31).

[Smi56] W. R. Smith. "Product Differentiation and Market Segmentation as Alternative Marketing Strategies". In: *Journal of Marketing* 21.1 (1956), pp. 3–8. DOI: 10.1177/002224295602100102 (p. 18).

[Spa] Sparkasse-Finanzportal GmbH. *Was ist Phishing?* Was ist Phishing? URL: https://www.sparkasse.de/service/sicherheit-im-internet/was-ist-phishing.html (accessed on 17.06.2021) (pp. 2, 12).

[Squ13] K. D. Squire. "Video Game–Based Learning: An Emerging Paradigm for Instruction". In: *Performance Improvement Quarterly* 26.1 (2013), pp. 101–130 (p. 15).

[SRV15] S. Stockhardt, B. Reinheimer, and M. Volkamer. "Über Die Wirksamkeit von Anti-Phishing-Training". In: *Mensch Und Computer 2015 – Workshopband*. Ed. by A. Weisbecker, M. Burmester, and A. Schmidt. Berlin: De Gruyter Oldenbourg, 2015, pp. 647–655 (pp. 169, 170).

[SS16] A. Streicher and J. D. Smeddinck. "Personalized and Adaptive Serious Games". In: *Entertainment Computing and Serious Games*. Ed. by R. Dörner, S. Göbel, M. Kickmeier-Rust, M. Masuch, and K. Zweig. Vol. 9970. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 332–377 (pp. 19, 20, 26, 29, 31).

[STZ04] K. Salen, K. S. Tekinbaş, and E. Zimmerman. *Rules of Play: Game Design Fundamentals*. MIT Press, 2004. 680 pp. (p. 13).

[Tal14] S. Talib. "Personalising Information Security Education". PhD thesis. Plymouth, UK: University of Plymouth, 2014. 385 pp. URL: http://hdl.handle.net/10026.1/2896 (accessed on 13.08.2021) (pp. 23–26, 31).

[TK14] P. Toth and P. Klein. "A Role-Based Model for Federal Information Technology/ Cyber Security Training". In: *NIST Special Publication* (2014), p. 154 (pp. 8, 9).

[Tli+19] A. Tlili et al. "Does Providing a Personalized Educational Game Based on Personality Matter? A Case Study". In: *IEEE Access* 7 (2019), pp. 119566–119575. DOI: 10.1109/ACCESS.2019.2936384 (p. 27).

[TMJ17] J.-N. Tioh, M. Mina, and D. W. Jacobson. "Cyber security training a survey of serious games in cyber security". In: *2017 IEEE Frontiers in Education Conference (FIE)*. Frontiers in Education. Indianapolis: IEEE, 2017, pp. 1–5. DOI: 10.1109/FIE.2017.8190712 (pp. 39, 40).

[Tse+11] S. Tseng, K. Chen, T. Lee, and J. Weng. "Automatic content generation for anti-phishing education game". In: *2011 International Conference on Electrical and Control Engineering*. International Conference on Electrical and Control Engineering. ICECENG '11. Yichang, China: IEEE, 2011, pp. 6390–6394. DOI: 10.1109/ICECENG.2011.6056921 (pp. 169, 170).

[Tse+15] S.-S. Tseng, T.-Y. Yang, J.-F. Weng, and Y.-J. Wang. "Building a game-based internet security learning system by ontology crystallization approach". In: *Proceedings of the International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government*. International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government. EEE '15. Las Vegas, NV, USA: CSREA Press, 2015, p. 6 (pp. 169, 170).

[TT20] D. Thornton and F. Turley. "Analysis of Player Behavior and EEG Readings in a Cybersecurity Game". In: *Proceedings of the 2020 ACM Southeast Conference*. ACM Southeast Conference. ACM SE '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 149–153. DOI: 10.1145/3374135.3385276 (pp. 25, 26, 31).

[vdAkk+06] J. van den Akker, K. Gravemeijer, S. McKenney, and N. Nieveen. *Educational Design Research*. Routledge, 2006. 177 pp. (p. 33).

[VF18] I. Vasileiou and S. Furnell. "Enhancing Security Education - Recognising Threshold Concepts and Other Influencing Factors:" in: *Proceedings of the 4th International Conference on Information Systems Security and Privacy*. International Conference on Information Systems Security and Privacy. ICISSP '18. Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications, 2018, pp. 398–403. DOI: 10.5220/0006646203980403 (pp. 23–25).

[VF19] I. Vasileiou and S. Furnell. "Personalising Security Education: Factors Influencing Individual Awareness and Compliance". In: *Information Systems Security and Privacy*. International Conference on Information Systems Security and Privacy. Ed. by P. Mori, S. Furnell, and O. Camp. Vol. 977. ICISSP '18. Cham: Springer International Publishing, 2019, pp. 189–200. DOI: 10.1007/978-3-030-25109-3_10 (pp. 23–25).

[Vol+17] M. Volkamer, K. Renaud, B. Reinheimer, and A. Kunz. "User Experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn". In: *Computers & Security* 71 (2017), pp. 100–113. DOI: 10.1016/j.cose.2017.02.004 (p. 12).

[VSB20] M. Volkamer, M. A. Sasse, and F. Boehm. "Analysing Simulated Phishing Campaigns for Staff". In: *Computer Security*. Ed. by I. Boureanu et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 312–328. DOI: 10.1007/978-3-030-66504-3_19 (pp. 2, 12, 105).

[Vuk12] E. Vuksani. "Device Dash: Designing, Implementing, and Evaluating an Educational Computer Security Game". Thesis. Wellesley, MA, USA: Wellesley College & MIT Lincoln Laboratory, 2012. 199 pp. URL: https://repository.wellesley.edu/object/ir356 (accessed on 04.01.2022) (pp. 169, 170).

[Wen+19] Z. A. Wen, Z. Lin, R. Chen, and E. Andersen. "What.Hack: Engaging Anti-Phishing Training Through a Role-Playing Phishing Simulation Game". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk). CHI '19. New York, NY, USA: Association for Computing Machinery, 2019. DOI: 10.1145/3290605.3300338 (pp. 2, 58, 169, 170).

[WH13] J. Wiemeyer and S. Hardy. "Serious Games and Motor Learning: Concepts, Evidence, Technology". In: *Serious Games and Virtual Worlds in Education, Professional Development, and Healthcare*. IGI Global, 2013, pp. 197–220 (p. 15).

[Whi12] N. Whitton. "Games-Based Learning". In: *Encyclopedia of the Sciences of Learning*. Ed. by N. M. Seel. Boston, MA: Springer US, 2012, pp. 1337–1340. DOI: 10.1007/978-1-4419-1428-6_437 (pp. 13, 15, 16).

[Wil+98] M. Wilson, D. E. de Zafra, S. I. Pitcher, J. D. Tressler, and J. B. Ippolito. *Information Technology Security Training Requirements: A Role- and Performance-Based Model*. NIST SP 800-16. Gaithersburg, MD: National Institute of Standards and Technology, 1998, p. 200. DOI: 10.6028/NIST.SP.800-16 (p. 8).

[WJZ18] P. Weanquoi, J. Johnson, and J. Zhang. "Using a Game to Improve Phishing Awareness". In: *Journal of Cybersecurity Education, Research and Practice* 20182/2.2 (2018), pp. 1–14 (pp. 58, 169, 170).

[WSZ20] Z. Wang, L. Sun, and H. Zhu. "Defining Social Engineering in Cybersecurity". In: *IEEE Access* 8 (2020), pp. 85094–85115. DOI: 10.1109/ACCESS.2020.2992807 (p. 11).

[Xia+18] Z. Xiao et al. "Cubicle: An Adaptive Educational Gaming Platform for Training Spatial Visualization Skills". In: *23rd International Conference on Intelligent User Interfaces*. International Conference on Intelligent User Interfaces. IUI '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 91–101. DOI: 10.1145/3172944.3172954 (p. 27).

[Xie+19] H. Xie, H.-C. Chu, G.-J. Hwang, and C.-C. Wang. "Trends and Development in Technology-Enhanced Adaptive/Personalized Learning: A Systematic Review of Journal Publications from 2007 to 2017". In: *Computers & Education* 140 (2019), p. 103599. DOI: 10.1016/j.compedu.2019.103599 (pp. 29, 30).

[YA14] E. O. Yeboah-Boateng and P. M. Amanor. "Phishing, SMiShing & Vishing: An Assessment of Threats against Mobile Devices". In: *Journal of Emerging Trends in Computing and Information Sciences* 5.4 (2014), pp. 297–307 (pp. 1, 11).

[Yan+12] C. Yang, S. Tseng, T. Lee, J. Weng, and K. Chen. "Building an Anti-phishing Game to Enhance Network Security Literacy Learning". In: *2012 IEEE 12th International Conference on Advanced Learning Technologies*. International Conference on Advanced Learning Technologies. Vol. 12. Rome, Italy: IEEE, 2012, pp. 121–123. DOI: 10.1109/ICALT.2012.174 (pp. 169, 170).

[Yas+18] A. Yasin, L. Liu, T. Li, J. Wang, and D. Zowghi. "Design and Preliminary Evaluation of a Cyber Security Requirements Education Game (SREG)". In: *Information and Software Technology* 95 (2018), pp. 179–200. DOI: 10.1016/j.infsof.2017.12.002 (pp. 169, 170).

[YNT08] W. D. Yu, S. Nargundkar, and N. Tiruthani. "A Phishing Vulnerability Analysis of Web Based Systems". In: *2008 IEEE Symposium on Computers and Communications*. IEEE Symposium on Computers and Communications. 2008, pp. 326–331. DOI: 10.1109/ISCC.2008.4625681 (pp. 1, 11).

B Auxiliary Materials

B.1 Publications on Games Data Set

Table 32: Summarized analysis results of POG data set (x = identified category; g = guessed category; y = yes; n = no; – = n/a)

	Target Group					Educational Context								
Reference	End-users	Non-CS Students	CS Students	Employees	IT employees	Primary School	Middle School	High School	College/University	Corporate	Informal	available?	focus: phishing?	digital?
[She+07]	x										x	y	y	y
[HGG15]	x										x	n	y	y
[Huy+17]		x							x			y	y	y
[WJZ18]			x						x			y	y	y
[GKG15]		x				x	x					y	n	y
[Lu18]			x						x			y	n	y
[KW18]				x						x		y	n	y
[BC14]	x										x	y	y	
[Wen+19]	x										x	y	y	y
[Gey19]	x										x	y	n	y
[ABP18]	x										x	n	n	y
[ALM15]	g										g	n	y	y
[BA19]	g										g	n	y	y
[BC16]	x										x	–	y	n
[Bau+17]		x				x	x					n	n	y
[BP16]					x					x		–	n	n
[BPF16]				x						x		–	n	n
[Bha19]	x										x	n	n	y
[CMB11]		x							x			n	n	y
[CJ+18]				x						x		n	y	y
[Con+07]			x						x			n	y	y
[FMS19]				x						x		n	n	y
[Fre+19]				x						x		–	n	n
[GP13]	x										x	–	n	n
[Heb+17]					x					x		n	y	y
[Kat+17]	g										g	n	n	y
[Kul19]	g										g	n	n	y
[Lop+18]		x					x					n	y	y
[MPJ18]		x				x	x					n	n	y
[MAB17]	x										x	n	y	y
[MvNvS10]				x						x		n	n	y
[Ola+14]		x						x				n	n	y
[OZ15]	g										g	n	n	y
[RL16]					x					x		–	n	n
[SRV15]	x										x	–	y	n
[Tse+11]	g										g	n	n	y
[Tse+15]	x	x					x				x	–	n	n
[Vuk12]	x										x	n	n	y
[Yan+12]		x							x		x	n	y	y
[Yas+18]			x						x			–	n	n

Table 33: Summarized analysis results of POG data set (x = identified category; g = guessed category; y = yes; n = no; – = n/a)

	factual						conceptual						procedural						meta-cognitive					
Reference	remember	understand	apply	analyze	evaluate	create	remember	understand	apply	analyze	evaluate	create	remember	understand	apply	analyze	evaluate	create	remember	understand	apply	analyze	evaluate	create
[She+07]	x						x	x					x		x									
[HGG15]	x	x	x				x	x	x				x	x	x									
[Huy+17]	x	x					x	x																
[WJZ18]	x	x					x																	
[GKG15]	x	x	x	x	x		x	x	x	x	x		x	x										
[Lu18]	x	x	x	x	x		x	x	x	x	x		x	x	x	x	x							
[KW18]	x	x					x	x	x						x									
[BC14]	x	x					x	x					x											
[Wen+19]	x	x					x	x																
[Gey19]	x	x					x	x					x	x										
[ABP18]	x	x					x	x																
[ALM15]	x						x																	
[BA19]	x						x																	
[BC16]	x	x	x				x	x	x				x	x	x									
[Bau+17]	x	x					x	x																
[BP16]		x						x	x	x				x	x	x	x	x	x	x	x	x		
[BPF16]	x	x	x				x	x	x				x	x										
[Bha19]	x	x					x	x																
[CMB11]	x												x		x									
[CJ+18]	x	x					x	x					x		x									
[Con+07]	x	x		x			x							x	x									
[FMS19]	x						x																	
[Fre+19]			x	x	x						x	x				x	x	x	x		x	x	x	
[GP13]		x	x	x	x		x		x					x		x	x				x			
[Heb+17]			x				x						x											
[Kat+17]							x								x									
[Kul19]	x	x					x	x																
[Lop+18]	x	x	x				x																	
[MPJ18]	x	x	x				x	x	x				x	x	x				x	x				
[MAB17]	x	x					x	x																
[MvNvS10]	x		x	x			x						x		x									
[Ola+14]	x	x	x	x			x						x		x									
[OZ15]	x	x	x				x	x																
[RL16]				x	x	x					x					x	x	x	x	x	x		x	x
[SRV15]	x			x			x	x		x														
[Tse+11]	x	x					x	x																
[Tse+15]	x						x																	
[Vuk12]	x	x					x	x																
[Yan+12]	x						x												x					
[Yas+18]	x	x	x	x		x			x				x	x			x		x		x			

B.2 Game Prototypes

B.2.1 Analysis Game

Gameplay

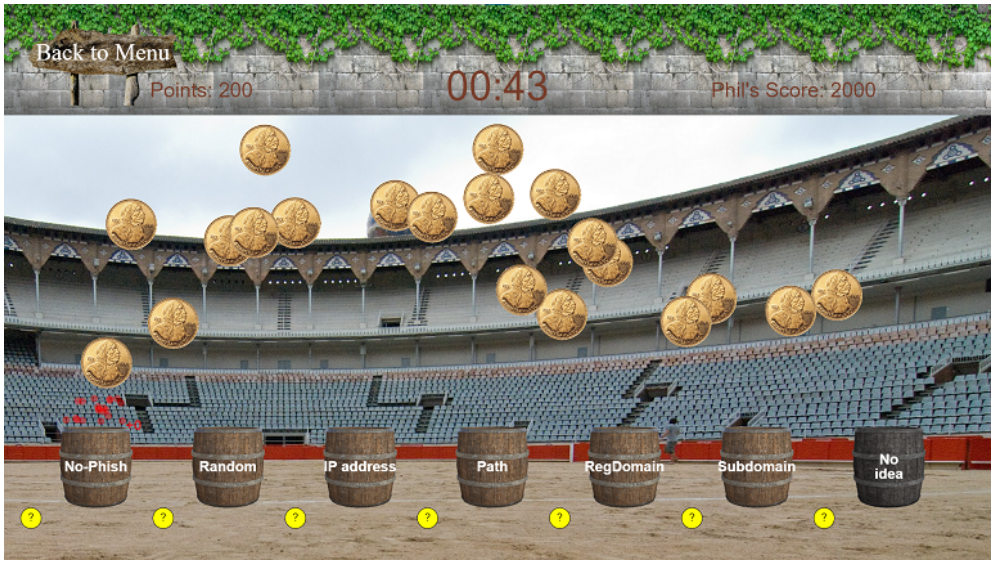

Figure 21: Screenshot of the analysis game showing a red aura above the bucket "No-Phish" indicating a recent incorrect classification.

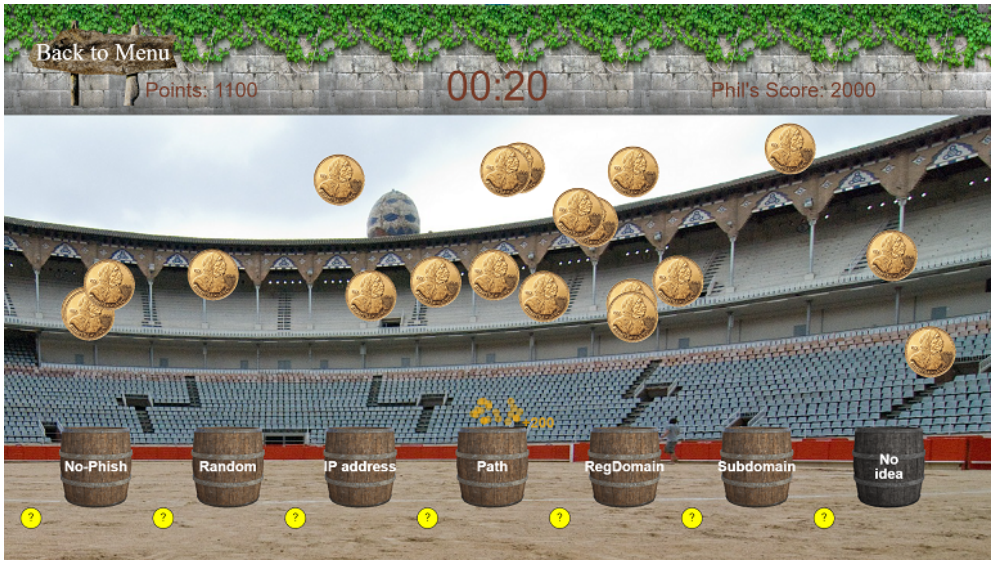

Figure 22: Screenshot of the analysis game showing a yellow aura above the bucket "Path" indicating a recent classification with correct tendency.

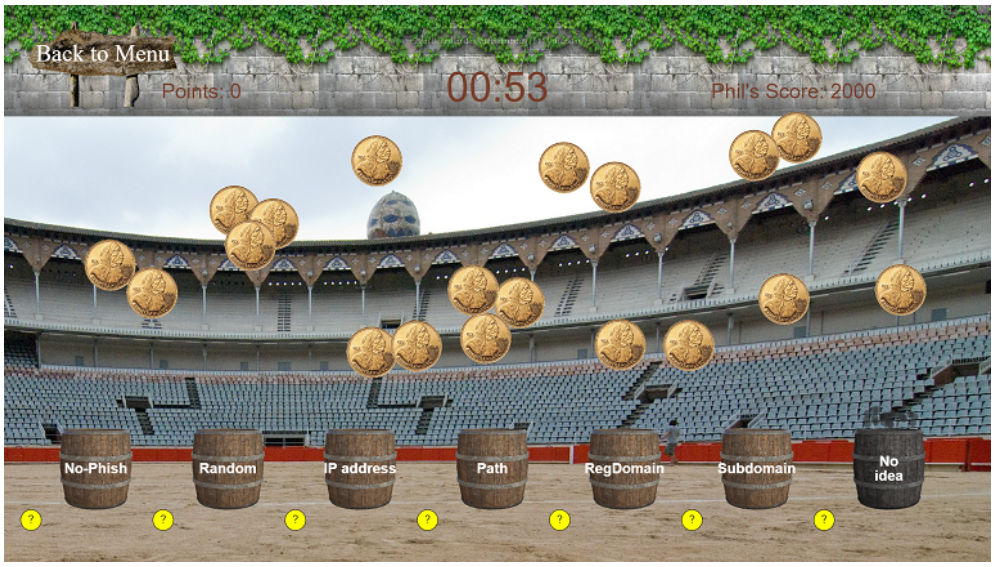

Figure 23: Screenshot of the analysis game showing a black aura above the bucket “No idea” indicating that players have discarded the last URL

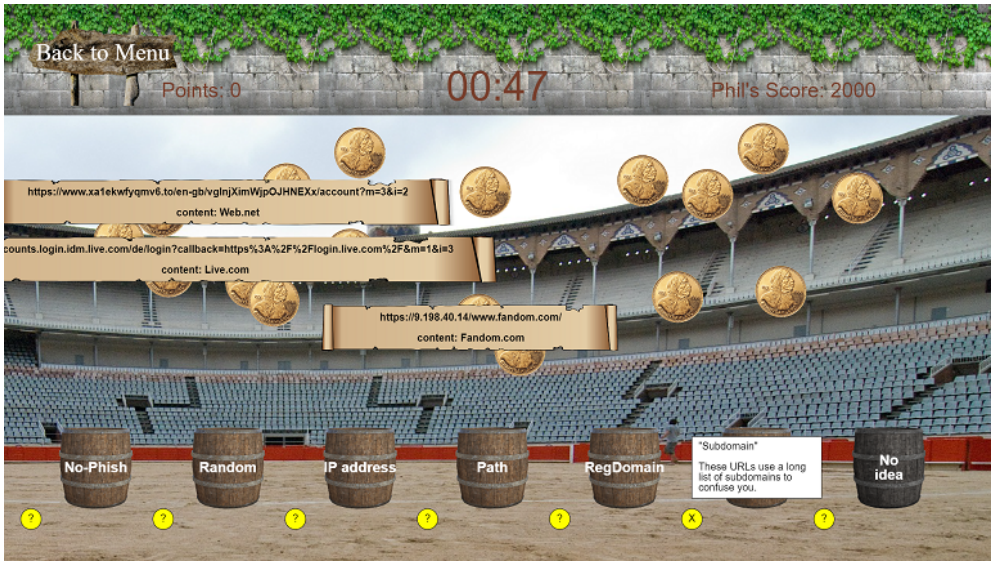

Figure 24: Screenshot of the analysis game showing an opened info box providing explanation of the URL category ‘Subdomain’.

Feedback

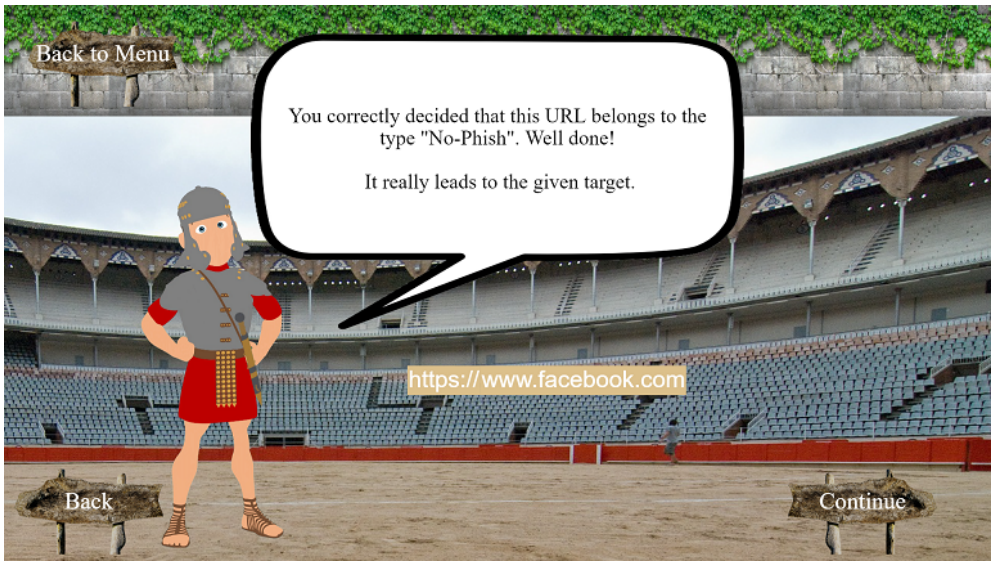

Figure 25: Feedback in the analysis game: Correct classification for URL of category 'No-Phish'.

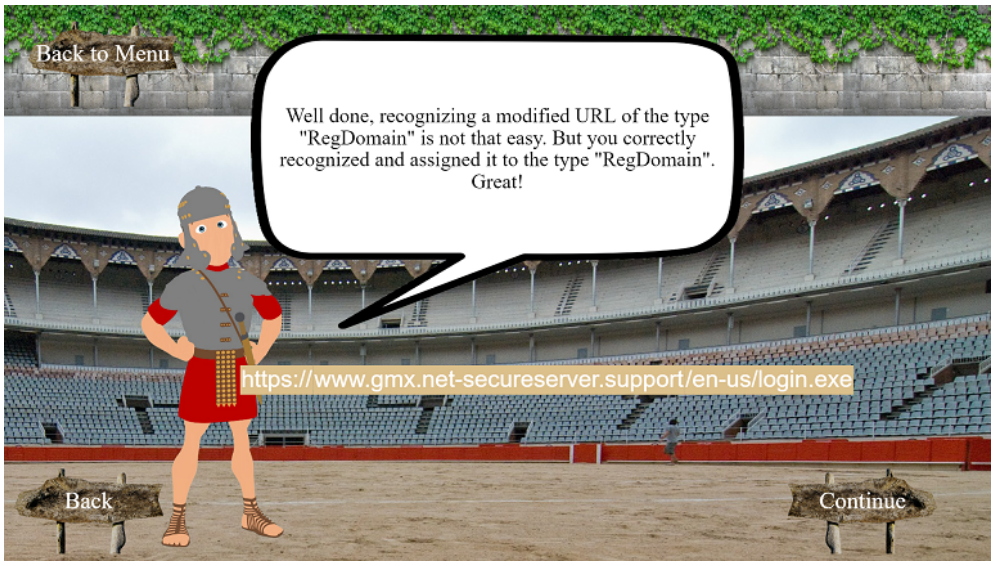

Figure 26: Feedback in the analysis game: Correct classification for URL of category 'RegDomain'.

Figure 27: Feedback in the analysis game: Incorrect classification for URL of category 'NoPhish'.

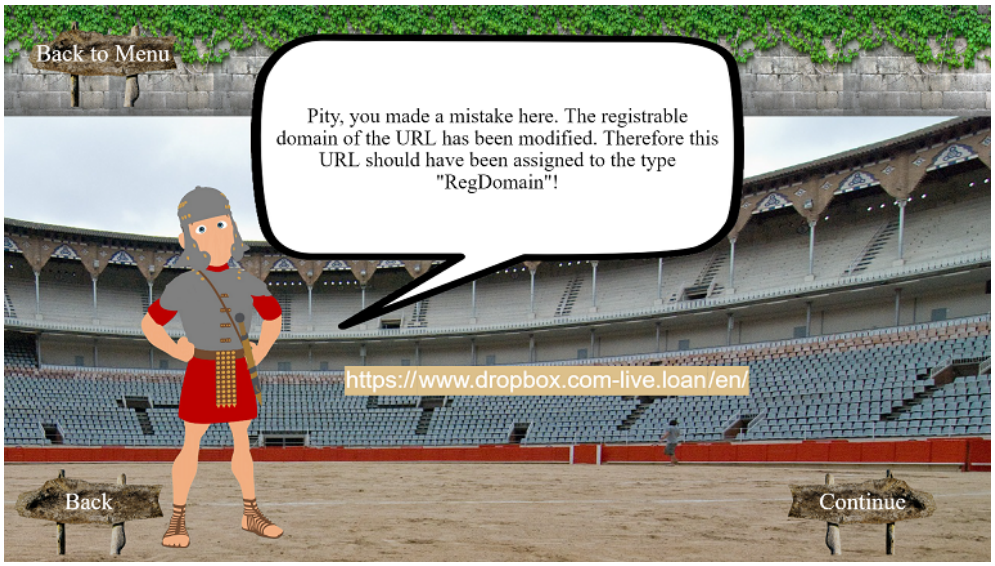

Figure 28: Feedback in the analysis game: Incorrect classification for URL of category 'RegDomain'.

Figure 29: Feedback in the analysis game: Discarded URL of category 'No-Phish' (similar feedback for other categories).

Figure 30: Feedback in the analysis game: Discarded URL of category 'RegDomain' (equal for other categories).

B.2.2 Creation Game

Gameplay

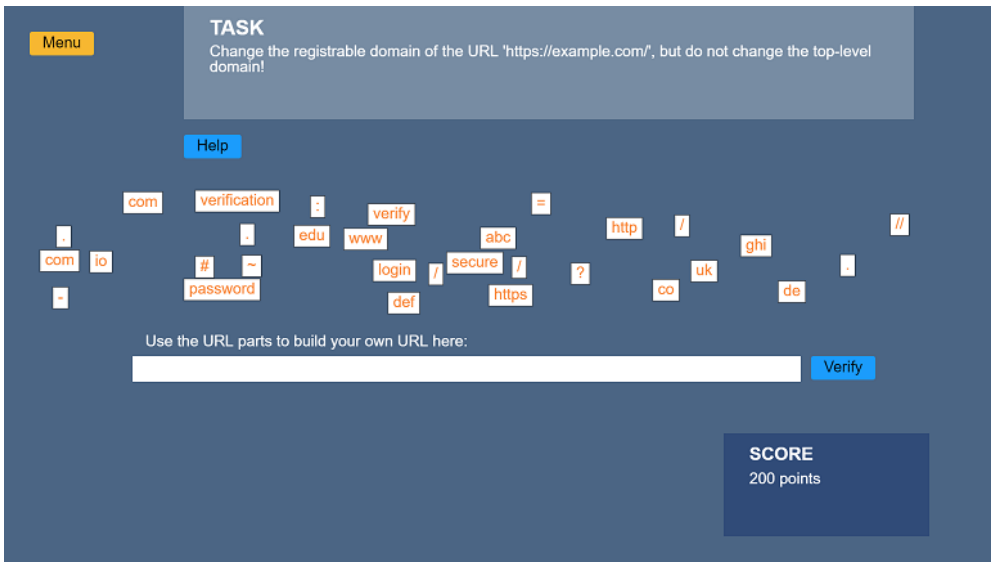

Figure 31: Screenshot of the creation game showing the initial state of the first level (second preset).

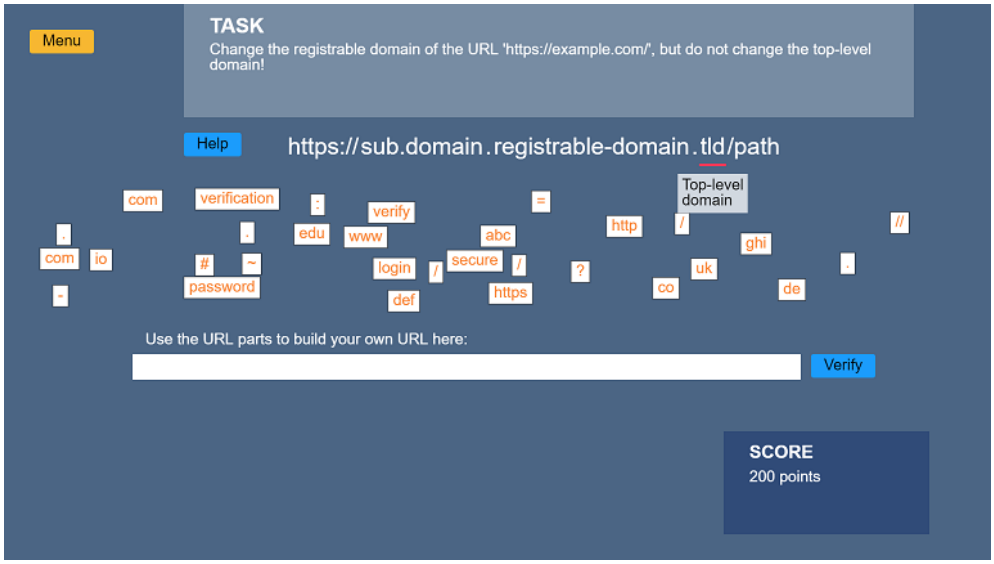

Figure 32: Screenshot of the creation game showing the helper element (after clicking on the 'Help' button).

Feedback

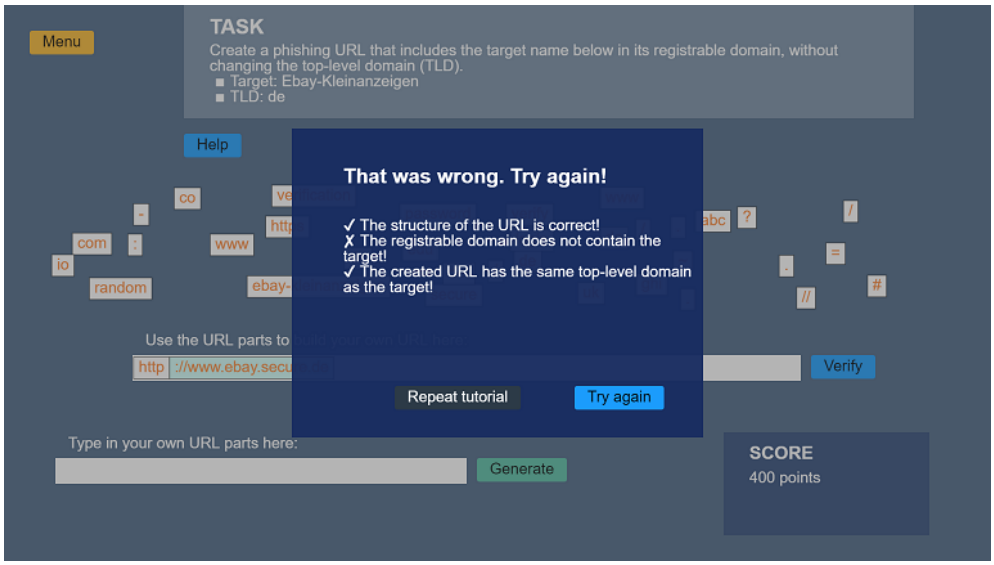

Figure 33: Feedback in the creation game: Fail check for missing target in RD.

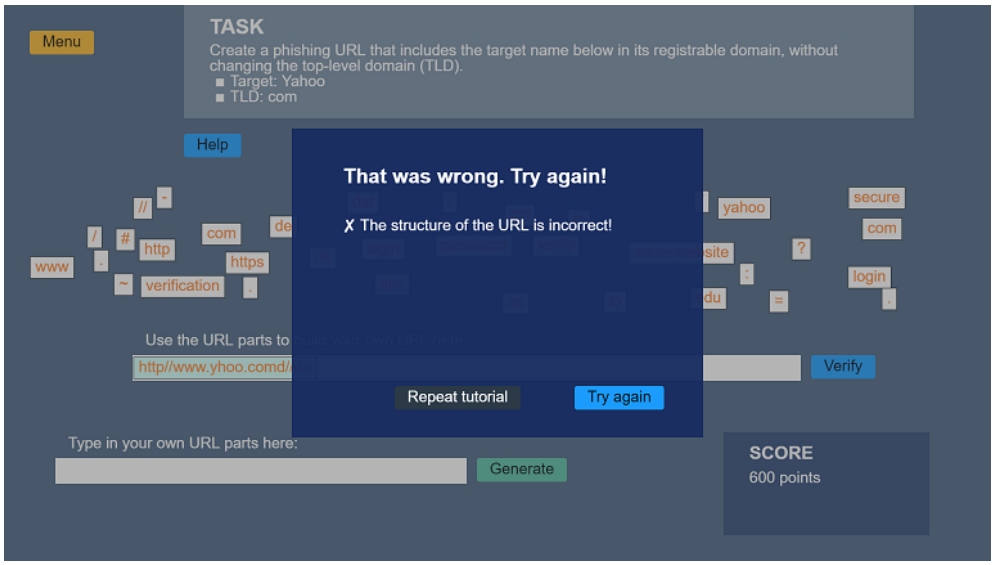

Figure 34: Feedback in the creation game: Incorrect URL structure.

B.2.3 Preliminary Evaluation

Instructions

Practical Session with ERBSE

If not already open in your browser, please open:

https://learntech.rwth-aachen.de/phishing-creation-game/

1. Play the game! (30 min)

Keep in mind (or write down) any problems or misunderstandings you may encounter.

2. Fill out the survey! (10 min)

Try to understand and answer all questions, ask if you are unclear about anything.

3. Evaluate the game! (15 min)

Get together in groups of two and discuss your thoughts about the games. Write down short notes on the last pages of the questionnaire:

- What you like or dislike about the game (Design, content, progression, etc.)?
- Specific problems with the game (things that did not work, were broken or unexpected)
- Lack of clarity (in content, task descriptions, feedback, ...)
- Misunderstandings, troubles in completing the given tasks
- What did you feel was still lacking? What has to be improved to make the game playable/more enjoyable?
- General feedback for this study and session

Figure 35: Screenshot of Instructions Handout for Preliminary Evaluation

Questionnaire

Table 34: Questionnaire for Preliminary Evaluation (second version of the handout contains URL to analysis game

ID	Questionnaire item
Q1	What did you like about the game?
Q2	What did you dislike about the game?
Q3	While playing, I encountered problems within the game...
Q4	While playing, I thought of ways to improve the game...
Q5	Things that cannot be left unsaid...

B.3 Personalization Pipeline

B.3.1 Selection Interface

Table 35: Services provided in the Selection Interface

Category	Service name	URL
Entertainment	Netflix	https://www.netflix.com/
	Twitch	https://www.twitch.tv/
	Fandom	https://www.fandom.com/
	Disney Plus	https://www.disneyplus.com/
	Artstation	https://www.artstation.com/
	Origin	https://www.origin.com/
	RuneScape	https://www.runescape.com/
	Drakensang	https://www.drakensang.com/
News	Spiegel	https://www.spiegel.de/
	Washington Times	https://www.washingtontimes.com/
	Zeit	https://www.zeit.de/
	ZDF	https://www.zdf.de/
	Süddeutsche Zeitung	https://www.sueddeutsche.de/
	Frankfurter Allgemeine	https://www.faz.net/
	Focus	https://www.focus.de/
	Rhein-Neckar-Zeitung	https://www.rnz.de/
	Legal Tribune Online	https://www.lto.de/
Shopping	Aliexpress	https://de.aliexpress.com/
	Amazon	https://www.amazon.de/
	Ebay	https://www.ebay.de/
	Ebay-Kleinanzeigen	https://www.ebay-kleinanzeigen.de/
	Mobile	https://www.mobile.de/
	Otto	https://www.otto.de/
	Idealo	https://www.idealo.de/
	Immobilienscout24	https://www.immobilienscout24.de/
Communication	Facebook	https://www.facebook.com/
	Twitter	https://twitter.com/
	Live	https://www.live.com/
	Yahoo	https://www.yahoo.com/
	Vk	https://vk.com/
	Web	https://web.de/
	GMX	https://www.gmx.net/
	Protonmail	https://mail.protonmail.com/
Finance	PayPal	https://www.paypal.com/
	Klarna	https://app.klarna.com/
	Skrill	https://www.skrill.com/
	Commerzbank	https://www.commerzbank.de/
	Deutsche Bank	https://www.deutsche-bank.de/
	Paysafecard	https://www.paysafecard.com/
	Union Bank	https://www.unionbank.com/
	Targobank	https://www.targobank.de/
Other	Google	https://www.google.com
	Apple	https://www.apple.com
	Wikipedia	https://de.wikipedia.org/
	Adobe	https://www.adobe.com/
	Dropbox	https://www.dropbox.com/
	iCloud	https://www.icloud.com/
	Slack	https://slack.com/
	Shopify	https://accounts.shopify.com/
	Codecademy	https://www.codecademy.com/
	Anki	https://ankiweb.net/

B.3.2 URL Generator Rule Set

Table 36: URL generator rule set for the creation of benign (b) or malicious (m) URLs.

Rule name	Type	Explanation	Example
loginRoute	b	Replaces path of original URL with Login-like path	example.de/login
addQueryParameter	b	Adds a random query parameter with key 'm' while keeping existing query parameters	example.de?m=9&i=4
addLanguagePath	b	Prepends a language path to the existing path	example.de/en-gb/login
subdomainLoginRoute	b	Adds a login route to subdomain	accounts.login.example.com/
trivialBenign	b	Returns original, unchanged input URL	example.de/
replaceHomoglyphsNormal	m	Replaces single characters with similar looking characters	exarnple.de/
hyphenSplitDomain	m	Adds a hyphen at random position, except beginning or end of the hostname	exarnple.de/
addSuspiciousKeyword	m	Adds a suspicious keyword separated by a hyphen to hostname	example-secure.de/
ipv4Simple	m	Replaces registrable domain by a random IP address and moves original hostname to the beginning of the path.	8.8.8.8/example.de/path
ipv4Hostname	m	Replaces registrable domain by a random IP address and uses original hostname including suspicious keyword as path	8.8.8.8/example-secure
omitVowel	m	Omits a random vowel from hostname, except from beginning or end of the hostname	exmple.de/
omitConsonant	m	Omits a random consonant from hostname, except from beginning or end of the hostname	exaple.de/
randomDomain	m	Replaces hostname by a totally random domain and moves original hostname to the beginning of the path	jadfhasdfasd.org/example.de/path
replaceVowel	m	Replaces a vowel in the hostname with another vowel	exampla.de/
replaceRandom	m	Replaces a random letter in the hostname with another random letter	examule.de/
subDomain	m	Moves the registrable domain into the subdomain of a random URL	example.de.kasj.de/
subDomainSuspicious	m	Adds a suspicious keyword to the hostname and moves registrable domain to subdomain of a random URL	example-secure.de.rajskdas.de/
swapLetterInside	m	Swaps two letters in the hostname except at the beginning or end of the hostname	exapmle.de/
swapLetterOutside	m	Swaps two letters at the beginning or end of the hostname	xeample.de/
totalRandom	m	Replaces registrable domain with random string	adfag3gfasfga.ru/
unfamiliarTld	m	Moves hostname to subdomain and creates a new registrable domain that starts with the original TLD following a hyphen, a deceptive keyword and an unfamiliar TLD.	example.de-secure.app/
urlEncodedAfterDomain	m	Moves registrable domain to subdomain and uses random hex encoded registrable domain	example.de@%23...

B.4 Comparative User Studies

B.4.1 Instructions

The following instructions were used to guide through the study session and follow the procedure described in Chapter 8.1.2 (see Listing 2 for a German translation).

Listing 1: Instructions for Briefing

Welcome to our user study. We are still waiting until the remaining participants join us in the Zoom Room.

So, now it looks like we are complete. Once again, a warm welcome. My name is **NAME1**, and together with my colleague **NAME2**, we are looking forward to your contribution to our study.

Is there a need to formulate the following instructions in German? Does everybody understand English fluently? If not, please tell us so that we can provide German instructions.

First, we would like to briefly explain the procedure of the study:

You are going to answer a two−part questionnaire in this study. The first part is an introductory questionnaire, and the second part is the closing questionnaire. Between part 1 and part 2, you will play a digital anti−phishing learning game. Please submit your results when you have reached the end of the questionnaire. We will then wait until everyone has finished and do a short final round in which we will be happy to answer questions about the study.

We would also like to clarify the requirements of the study again:

It is important that you use either the Google Chrome or Mozilla Firefox browser on a computer or notebook for the questionnaire and the learning game. Mobile devices are not fully supported. Please also ensure that your device has sufficient battery power and that you are not disrupted during the study.

When completing the questionnaire, we ask you to answer honestly and spontaneously. Please also do not try to research the answers to questions during the study. This is about your judgment and opinion. Do not spend too much time on one question, but continue working on the questionnaire.

If you encounter, e.g., a technical problem while working on the questionnaire or during the game, please let us know by speaking up. We will then help you in a secondary room in Zoom so that you can continue.

At the beginning of the study, my colleague will give you a short verbal introduction to phishing and an example.

Handing over to colleague for phishing definition and example

    Phishing is an internet−based act of deception whereby impersonation is used to obtain information from a target.

    Here is a common example:

        Mr. Smith receives an email from a popular online shop that urges him to click on a link to verify his customer information.

        He complies with this request and follows the link to a website of the online shop, where he enters his password and further personal information.

        Several weeks later, he notices orders at the online shop that he does not remember placing.

17 The email and website were fake, Mr. Smith has become the victim of a Phishing attack.
18 If Mr. Smith had taken a closer look at the fake website, he might have noticed differences to the real website of the online store.
19 This study is about anti−phishing learning games. These are intended to teach users how to recognize such differences between the real and fake websites.
20 Are there any questions before we start?
21 **Handing over to colleague for last instructions**
22 We will now send you the link to the questionnaire in the Zoom chat. You have received an access key in the email yesterday. Attention! Please do not confuse it with the password to Zoom. Please enter the access key on the website to start the questionnaire.
23 And again, thank you for taking the time to participate in the study. Now, it would be really important for our research that you answer honestly and work through the study completely. Your data will be evaluated according to data protection regulations and cannot be traced back to you.
24 **Copy & paste questionnaire link to Zoom chat**
25 Now you will find the link to the survey in the Chat in Zoom.
26 Note that the default language for the questionnaire and learning game is German. You can change this to English in the settings menu.

Listing 2: German Translation of Instructions for Briefing

1 Herzlich willkommen zu unserer Nutzerstudie. Wir warten noch bis die anderen Teilnehmerinnen und Teilnehmer auch in den Zoom−Raum eingetreten sind.
2 So, nun scheinen wir komplett zu sein. Nochmals herzlich willkommen. Mein Name ist **NAME1** und gemeinsam mit meinem Kollegen **NAME2** freuen wir uns auf Ihre Teilnahme an unserer Studie.
3 Gibt es den Bedarf, die nachfolgenden Anweisungen auf Englisch zu formulieren? Does everybody understand German fluently? If not, please tell us, so we can provide English instructions.
4 Zuerst möchten wir Ihnen den Ablauf der Studie kurz erläutern: Sie werden im Rahmen dieser Studie einen Fragebogen mit zwei Teilen bearbeiten. Der erste Teil ist ein Einstiegsfragebogen und der zweite Teil ist der Abschlussfragebogen. Zwischen Teil 1 und Teil 2 werden Sie ein digitales Anti−Phishing Lernspiel spielen. Wenn Sie am Ende des Abschlussfragebogens angekommen sind, können Sie Ihre Ergebnisse abschicken. Wir würden dann warten bis alle fertig sind und eine kurze Abschlussrunde machen, in der wir gerne Fragen zur Studie beantworten.
5 Wir möchten zudem noch einmal kurz die Bearbeitungsbedingungen abklären: Wichtig, ist, dass Sie für die Umfrage und das Lernspiel entweder den Browser Google Chrome oder Mozilla Firefox auf einem Computer oder Notebook benutzen. Mobilgeräte werden nicht vollständig unterstützt. Stellen Sie bitte auch sicher, dass ihr Gerät ausreichend Strom besitzt und Sie nicht während der Studie unterbrochen werden.
6 Bei der Bearbeitung des Fragebogens möchten wir Sie bitten ehrlich und aus dem Bauch heraus zu antworten. Bitte versuchen Sie auch nicht während der Studie die Antworten auf Fragen zu recherchieren. Es geht hierbei um Ihre Einschätzung und Meinung. Halten Sie sich nicht zu lange an einer Frage auf, sondern bearbeiten Sie weiter den Fragebogen.
7 Wenn Sie auf ein z. B. technisches Problem in der Bearbeitung des Fragebogens oder während des Spiels treffen, melden Sie sich bitte kurz zu Wort. Wir werden Ihnen dann in einem Nebenraum in Zoom helfen, sodass Sie fortfahren können.
8 Zu Beginn der Studie wird mein Kollege Ihnen eine kurze mündliche Einführung zu Phishing sowie ein Beispiel vorstellen.
9 **Übergabe an Kollegen für Phishing-Definition und Beispiel**
10 Phishing ist eine internetbasierte Art von Betrug, in dem Täuschung verwendet wird, um die Informationen eines Opfers zu stehlen.
11 Nun folgt ein typisches Beispiel:
12 Herr Müller bekommt eine E−Mail von einem beliebten Online−Shop, in der er aufgefordert wird auf einen Link zu klicken, um seine Kundeninformationen zu überprüfen.
13 Er kommt dieser Aufforderung nach, wird auf eine Website des Online−Shops weitergeleitet und gibt dort, wie aufgefordert, sein Passwort und weitere persönlichen Daten ein.
14 Einige Wochen später stellt er fest, dass Bestellungen im Online−Shop auftauchen, an die er sich nicht erinnern kann.
15 Die E−Mail und Website waren gefälscht, Herr Müller ist Opfer eines Phishing Angriffs geworden.
16 Hätte Herr Müller sich die gefälschte Webseite genauer angeschaut, hätte er möglicherweise Unterschiede zur echten Webseite des Online−Shops bemerkt.

17 In dieser Studie geht es um Anti−Phishing Lernspiele. Diese sollen beibringen, wie man solche Unterschiede zwischen der echten und gefälschten Webseite erkennt.
18 Gibt es noch Fragen bevor wir loslegen?
19 **Übergabe an Kollegen für letzte Anweisungen**
20 Wir werden Ihnen jetzt den Link zur Umfrage im Chat von Zoom schicken. Sie haben in der E−Mail gestern einen Zugangsschlüssel erhalten (Achtung! Bitte nicht mit dem Passwort zu Zoom verwechseln). Bitte geben Sie den Zugangsschlüssel zum Starten des Fragebogens auf der Webseite ein.
21 Und nochmal eine Bitte: Sie haben sich dankenswerterweise die Zeit für die Studienteilnahme genommen. Nun wäre es für unsere Forschung wirklich wichtig, dass Sie ehrlich antworten und die Studie vollständig bearbeiten. Ihre Daten werden datenschutzkonform ausgewertet und sind nicht auf Sie zurückführbar.
22 **Kopieren & Einfügen des Fragebogen-Links in den Zoom-Chat**
23 Nun finden Sie im Chat in Zoom den Link zur Umfrage.
24 Sie können nun mit der Bearbeitung starten. Bitte beachten Sie, dass das Spiel ein explizites Ende hat und sie anschließend zurück zum Fragebogen wechseln können um dort den zweiten Teil zu bearbeiten.
25 Bei Fragen melden Sie sich bitte kurz zu Wort, sodass wir Ihnen helfen können.

B.4.2 Questionnaires and Additional Results

Table 37: German translation of the **Perception of Phishing** questionnaire with three constructs: perceived susceptibility (PSU), perceived severity (PSE) and perceived threat (PTH).

Item	Item text
PSU1	Es ist sehr wahrscheinlich, dass ich Opfer eines Phishing-Angriffs werden kann.
PSU2	Meine Chancen, Opfer eines Phishing-Angriffs zu werden, sind groß.
PSU3	Ich denke nicht, dass ein Phishing-Angriff auf mich erfolgreich sein wird.
PSE1	Ein Phishing-Angriff würde ohne mein Wissen persönliche Daten von meinem Gerät stehlen.
PSE2	Ein Phishing-Angriff stellt ein Risiko für meine Privatsphäre dar.
PSE3	Ich denke, ein Phishing-Angriff würde meine persönlichen Daten nicht ohne mein Wissen von meinem Gerät stehlen.
PSE4	Ich denke, ein Phishing-Angriff würde meine Privatsphäre nicht gefährden.
PTH1	Phishing-Angriffe stellen für mich eine Gefahr dar.
PTH2	Ein Phishing-Angriff ist eine Gefahr für meine persönlichen Daten.
PTH3	Es ist riskant, mein Gerät zu nutzen, wenn ich Ziel eines Phishing-Angriffs bin.
PTH4	Ich denke, dass ein Phishing-Angriff mir keinen Schaden zufügen wird.

Table 38: **Demographics** questionnaire (see Table 39 for a German translation)

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 39: German translation of the **Demographics** questionnaire including answer types and options

Question	Answer type	Answer options
Ihr Geschlecht:	single-choice	Weiblich; Männlich; Divers; Keine Angabe
Wie alt sind Sie?	single-choice	14 oder jünger younger; 15-19; 20-24; 25-29; 30-34; 35-39; 40 oder älter; Keine Angabe
Was ist Ihr höchster Abschluss?	single-choice	Kein Schulabschluss; Haupt-/Realschule (Mittlere Reife); (Fach-)Abitur; Berufsausbildung; Bachelor; Master; Diplom; Promotion; Sonstiges; Keine Angabe
Haben Sie an Informatikunterricht (z.B. in der Schule oder Hochschule) teilgenommen?	single-choice	Kein Informatikunterricht; Weniger als 6 Monate; 6 - 12 Monate; 1 - 2 Jahre; Mehr als 2 Jahre; Keine Angabe
Wie schätzen Sie ihr Vorwissen in den folgenden Themen ein? (Informatik | IT-Sicherheit | Phishing)	6-point Likert scale	Keins; Sehr wenig; Wenig; Etwas; Viel; Sehr viel

Table 40: German translation of the **Behavioral Change Questionnaire** used in longitudinal testing with the four constructs: Application (App), Interest (Int), Behavior Change (BC), and Perceived Threat (PT).

Item	Item text
App1	Ich habe das was ich im Spiel gelernt habe in den letzten Monaten angewendet.
App2	Seit dem Spielen des Lernspiels überprüfe ich die URLs von Webseiten, bevor ich sie anklicke.
App3	Seit dem Spielen des Lernspiels überprüfe ich die URLs von Webseiten, bevor ich persönliche Daten eingebe (z. B. Zugangsdaten).
Int1	Das Spielen des Lernspiels hat mein Interesse an Phishing oder anderen IT-Sicherheitsthemen geweckt.
Int2	Ich würde gerne mehr über Phishing- oder weitere IT-Sicherheitsthemen lernen, indem ich Lernspiele spiele.
BC1	Seit dem Spielen des Lernspiels bin ich aufmerksamer gegenüber Phishing-Angriffen geworden.
BC2	Nach dem Spielen des Lernspiels habe ich mein Verhalten im Umgang mit URLs angepasst.
PT1	Nach dem Spielen des Lernspiels habe ich das Gefühl mich vor Phishing-Angriffen schützen zu können.
PT2	Nach dem Spielen des Lernspiels habe ich das Gefühl, dass es weniger wahrscheinlich ist, dass ich auf Phishing-Angriffe hereinfalle.

Table 41: 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 42: Absolute (and relative) results of the **Familiarity with 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

[a]Services were blinded so as not to reveal our country of origin.

Table 43: 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

C List of Figures

D List of Tables

E List of Listings

F List of Abbreviations

AI Artificial Intelligence 26, 28, 30
ANOVA Analysis of Variance 38, 99, 101, 102, 113, 114, 116, 123–125
API Application Programming Interface 75
APWG Anti-Phishing Working Group 11, 94

BRT Bloom's Revised Taxonomy 35, 36, 50–55, 59, 62, 63, 66, 109, 130, 134, 136, 141, 143

CS Computer Science 2, 9, 10, 25, 40, 43, 44, 47, 48, 51–53, 59, 75, 76, 97, 98, 113, 129
CTF Capture the Flag (competition) 25, 42

DC Developer Content 28, 29
DDA Dynamic Difficulty Adjustment 25, 26, 28, 31

EEG Electroencephalography 25

FBI Federal Bureau of Investigation 49

GA Game Analytics 18
GLA Game Learning Analytics 18

HCI Human Computer Interaction 19
HTML Hypertext Markup Language 18

ID Identifier 75
IDN Internationalized Domain Name 56, 58, 60
IP Internet Protocol (address) 60, 65, 69–71, 106, 117
ISA Information Security Awareness 8
ISE Information Security Education 8
IST Information Security Training 8
IT Information Technology 2, 3, 7–11, 15, 39–41, 43, 47, 49, 51–54, 59, 61, 97, 98, 113, 129, 130, 139, 143, 144
ITS Intelligent Tutoring System 26

JS JavaScript 18, 36

LA Learning Analytics 18

MFA Multi-factor Authentication 56, 58
MRQ Main Research Question 4, 144
MTLG Multi-Touch Learning Game (framework) 17, 18, 36, 74, 75, 86

NGLOB Narrative Game-based Learning Object 26
NPC Non-Player Character 14, 68

PC Procedural Content 28, 29
PCG Procedural Content Generation 28, 29, 32
PISE Personalising Information Security Education (framework) 24, 31

POG Publications on Games (data set) 52, 53, 56, 57, 59, 61, 63, 133

RD Registrable Domain 57, 65, 66, 69, 70, 72, 74, 177
RQ Research Question 92, 95, 97, 100–108, 110, 111, 115, 116, 118–121, 123, 125–127
RWTH Rheinisch-Westfälische Technische Hochschule 17, 34, 75, 98, 106

SMS Short Message Service 11, 56, 88
SRQ Sub Research Question 4, 142–144

TLD Top-Level Domain 65, 66, 70, 74, 94, 106

U2F Universal Second Factor 1, 12
UGC User-generated Content 28, 29
URL Uniform Resource Locator 1–5, 11–13, 16, 35, 37, 38, 54, 56–61, 63–74, 76–79, 82, 85–89, 91–97, 99–122, 124, 125, 127, 130–139, 141–145, 172–175, 177

VR Virtual Reality 15–17, 31

xAPI Experience API 75

G Statement of Originality

This dissertation is based on research and work that was conducted at the *Learning Technologies Research Group*, directed by Prof. Ulrik Schroeder. It was part of a cooperation with the *IT-Security Research Group*, directed by Prof. Ulrike Meyer, and the research training group "Human Centered Systems Security", sponsored by the state of North Rhine-Westphalia, Germany. The approaches and ideas presented in this thesis have strongly benefited from collaborative work and valuable discussions with colleagues from both institutions and the research training group. In his role as doctoral supervisor, Ulrik Schroeder contributed valuable feedback to all publications and helped to improve the writing and the overall research process. Similarly, Ulrike Meyer and Martin Wolf provided valuable feedback and guidance as part of the project ERBSE in the research training group. Further details on the contributions to individual publications is presented in the following.

[A] "Konzeption und Entwicklung eines interaktiven E-Mail-Interface für Anti-Phishing Lernspiele" by *D. Bayrak, R. Röpke, and U. Schroeder*

This paper presents an interactive email interface for anti-phishing learning games. The work originated from the student thesis project by Duygu Bayrak. My contribution to this paper was guidance throughout the student thesis project and detailed feedback and improvements to the paper.

[B] "Analyzing and Creating Malicious URLs: A Comparative Study on Anti-Phishing Learning Games" by *V. Drury, R. Roepke, U. Schroeder, and U. Meyer*

This paper presents a comparative evaluation of two game prototypes and a baseline implementation. It is a collaborative work with Vincent Drury, in which we both contributed equally. As such, designing and conducting the study as well as analyzing and discussing its results was done collaboratively.

[C] "Through a Mirror Darkly – On the Obscurity of Teaching Goals in Game-Based Learning in IT Security" by *K. Köhler, R. Röpke, and M. Wolf*

This paper analyzes learning outcomes in learning games for security education. My contribution to this paper is the data set of learning games for security education and its presentation in the paper. Klemens Köhler used this data set to perform different analysis and review steps to identify learning outcomes and explore whether games cover the complete pedagogical spectrum according to Bloom's Revised Taxonomy.

[D] ***Evaluation Data of Anti-Phishing Learning Games*** by *R. Röpke and V. Drury*

This repository contains the collected research data, anonymized for publication. It was collected together with Vincent Drury when conducting the comparative user studies described in this dissertation.

[E] **"A Pond Full of Phishing Games - Analysis of Learning Games for Anti-Phishing Education"** by *R. Roepke, K. Koehler, V. Drury, U. Schroeder, M. Wolf, and U. Meyer*

This paper summarizes the state-of-the-art of anti-phishing learning games. It is a collaborative work with Vincent Drury and Klemens Köhler, in which we all contributed equally. My main contribution was the literature research methodology and analysis of target groups and educational contexts. Klemens Köhler contributed the analysis of learning goals and game mechanics, while Vincent Drury analyzed the learning content of available learning games.

[F] **"Towards Personalized Game-Based Learning in Anti-Phishing Education"** by *R. Roepke, U. Schroeder, V. Drury, and U. Meyer*

This paper presents conceptual work regarding the personalization pipeline. My main contribution to this paper is the development and presentation of the concept with its three steps as well as related work. Vincent Drury contributed to the development of the concept in valuable discussions as well as the formalization and presentation in this paper.

[G] **"Exploring Different Game Mechanics for Anti-phishing Learning Games"** by *R. Roepke, V. Drury, U. Meyer, and U. Schroeder*

This paper presents the two game prototypes developed as part of this dissertation. My main contribution to this paper is the design and implementation of both prototypes in close cooperation with Vincent Drury. I took the lead in writing the main content of this paper, while Vincent Drury supported through critical review and improvements of the overall presentation of both games.

[H] **"A Modular Architecture for Personalized Learning Content in Anti-Phishing Learning Games"** by *R. Roepke, V. Drury, U. Schroeder, and U. Meyer*

This paper presents the implementation of the personalization pipeline. My main contribution to this paper is the implementation design and presentation with its different modules and components. I took the lead in writing the main content of this paper. Vincent Drury contributed to the implementation in various development sessions and supported to describe it in this paper.

[I] "Better the Phish You Know: Evaluating Personalization in Anti-Phishing Learning Games" by *R. Roepke, V. Drury, U. Meyer, and U. Schroeder*

This paper presents a comparative evaluation of a game and its personalized version after applying the developed personalization pipeline. It is a collaborative work with Vincent Drury, in which we both contributed equally. The paper reports on the design and conduction of the study as well the analysis and discussion of results and it was written collaboratively.

[J] "Exploring and Evaluating Different Game Mechanics for Anti-Phishing Learning Games" by *R. Roepke, V. Drury, U. Meyer, and U. Schroeder*

This paper presents a comparative evaluation of two new anti-phishing learning games. It is a collaborative work with Vincent Drury, in which we both contributed equally. The paper reports on the design and conduction of the study as well the analysis and discussion of results and it was written collaboratively.

[K] "More Than Meets the Eye - An Anti-Phishing Learning Game with a Focus on Phishing Emails" by *R. Roepke, V. Drury, P. Peess, T. Johnen, U. Meyer, and U. Schroeder*

This paper presents a game prototype of an anti-phishing learning game on the topic of phishing emails. It is a collaborative work with Vincent Drury, Philipp Peeß and Tobias Johnen. The paper reports on the design and implementation of the game prototype, including personalized level generation and first evaluation results in preliminary user studies. It was written collaboratively.

[L] "Mit der Lupe unterwegs - eine spiel-basierte Lernanwendung zu Sicherheit im Internet" by *R. Röpke, K. Larisch, S. Schöbel, and U. Schroeder*

This paper presents a game prototype developed in a supervised student thesis project by Sven Schoebel. While I initiated the projects and took lead in writing the main content of the paper, Kathrin Larisch contributed with valuable feedback and discussions throughout the design and development of the game. Sven Schöbel contributed the prototype as a result of his thesis and prepared the prototype for the presentation at the conference.

[M] "Spielbasierte Lernanwendungen für sicheren Umgang mit IT-Systemen und dem Internet" by *R. Röpke*

This paper presents the early stages of this dissertation project in which neither the concept of personalization nor any game prototypes were decided yet. It was presented at a doctoral colloquium as part of DELFI 2019.

[N] "Spielerisch sichere Teilhabe: Ein Review spiel-basierter Lernanwendungen über IT-Sicherheit und Sicherheitspraktiken" by *R. Röpke*

This paper presents a German translation of the review work presented in [RS19b]. It was presented at the "Junges Forum für Medien und Hochschulentwicklung", a forum for young doctoral researchers in their first years of research.

[O] "Teaching Defence Against the Dark Arts Using Game-Based Learning" by *R. Roepke and U. Schroeder*

This paper summarizes the state-of-the-art of learning games for security education as well as the classification of existing games using a classification model. My main contribution to this paper lies in the literature research and application of review steps and the classification model. This paper is an extended version of [P].

[P] "The Problem with Teaching Defence against the Dark Arts: A Review of Game-based Learning Applications and Serious Games for Cyber Security Education" by *R. Roepke and U. Schroeder*

This paper summarizes the state-of-the-art of learning games for security education. My main contribution to this paper lies in the literature research and application of review steps to identify the current state-of-the-art of existing learning games and their shortcomings.

[Q] "Mit G/P/S durch die Welt der spielbasierten Lernanwendungen und Serious Games für IT-Sicherheit" by *R. Röpke and U. Schroeder*

This paper presents a poster of the application of a classification model to a previously identified set of learning games for security education. My main contribution to this paper is the application of the classification and analysis of results to present it on the poster.

[R] "Phishing Academy: Evaluation of a Digital Educational Game on URLs and Phishing" by *S. Schoebel, R. Roepke, and U. Schroeder*

This paper presents a user study evaluating the anti-phishing learning game Phishing Academy. The work originated from the student thesis project by Sven Schoebel. My contribution to this paper was guidance throughout the student thesis project and detailed feedback and improvements to the paper.