TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
49/2020

Evaluating Cybersecurity-Related Competences through Simulation Exercises

STEN MÄSES

TAL TECH PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Department of Software Science
The dissertation was accepted for the defence of the degree of Doctor of Philosophy in Computer Science (Cybersecurity) 30/10/2020

Supervisor: Prof. Dr. Olaf Manuel Maennel,
Department of Software Science, School of Information Technologies,
Tallinn University of Technology
Tallinn, Estonia

Co-supervisors: Dr. Liina Randmann,
Department of Business Administration, School of Business and Governance
Tallinn University of Technology
Tallinn, Estonia

Prof. Dr. Stefan Sütterlin,
Faculty of Health and Welfare Sciences
Ostfold University College
B R A Veien 4, Halden, Norway

Opponents: Associate Professor Nickolas Falkner,
University of Adelaide,
Australia

Associate Professor Petri Ihantola,
University of Helsinki,
Finland

Defence of the thesis: 16/12/2020, Tallinn

Declaration:
Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology, has not been submitted for any academic degree elsewhere.

Sten Mäses

signature

European Union
European Regional
Development Fund

Investing
in your future

ISSN 2585-6898 (publication)
ISBN 978-9949-83-621-5 (publication)
ISSN 2585-6901 (PDF)
ISBN 978-9949-83-622-2 (PDF)
Printed by Koopia Niini & Rauam

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
49/2020

Küberturbe-alaste kompetentside hindamine simulatsiooniharjutuste abil

STEN MÄSES

TAL
TECH
KIRJASTUS

Contents

List of publications

The list of publications is ordered by the logical narrative—i.e. the suggested order to read them. Additionally, the extended abstract of this thesis has been published [56] but not separately mentioned in the current publication list.

I S. Mäses, B. Hallaq, and O. Maennel. Obtaining better metrics for complex serious games within virtualised simulation environments. In *ECGBL 2017 11th European Conference on Game-Based Learning*, pages 428–434. Academic Conferences and publishing limited, 2017

II S. Mäses, L. Randmann, O. Maennel, and B. Lorenz. Stenmap: Framework for evaluating cybersecurity-related skills based on computer simulations. In *Learning and Collaboration Technologies. Learning and Teaching*, pages 1–13. Springer, 2018

III M. Ernits, K. Maennel, S. Mäses, T. Lepik, and O. Maennel. From simple scoring towards a meaningful interpretation of learning in cybersecurity exercises. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security: ICCWS 2020*. Academic Conferences and publishing limited, 2020

IV S. Mäses, K. Kikerpill, K. Jüristo, and O. Maennel. Mixed methods research approach and experimental procedure for measuring human factors in cybersecurity using phishing simulations. In *ECRM 2019 18th European Conference on Research Methodology for Business and Management Studies*, pages 428–434. Academic Conferences and publishing limited, 2017

V S. Yari, S. Mäses, and O. Maennel. A method for teaching open source intelligence (OSINT) using personalised cloud-based exercises. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security: ICCWS 2020*. Academic Conferences and publishing limited, 2020

VI S. Mäses, H. Aitsam, and L. Randmann. A method for adding cyberethical behaviour measurements to computer science homework assignments. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, page 18. ACM, 2019

VII S. Mäses, O. Maennel, and S. Sütterlin. Using competency mapping for skill assessment in an introductory cybersecurity course. In *Educating Engineers for Future Industrial Revolutions - Proceedings of the 23rd International Conference on Interactive Collaborative Learning (ICL2020)*. Springer, 2020

Author's contributions to the publications

This chapter explains the contribution of the author of this thesis to the listed publications.

I As the main and leading author of this publication, the author analysed the main relevant performance metrics for measuring the skills of participants of virtual machine based serious games. The author proposed a theoretical concept how skills can be linked to different performance measurement categories, prepared the figures, and wrote the manuscript.

II As the main and leading author of this publication, the author conducted the case study, designed the processes, analysed the results, prepared the figures, and wrote the manuscript.

III As the co-author of this publication, the author contributed to the model creation, created the figures, and contributed to writing and revising the manuscript.

IV As the main and leading author of this publication, the author designed the method, participated in carrying out the technical experiments, and wrote the manuscript with the help of co-authors.

V As the co-author of this publication, the author contributed to the conduct of the case study, and wrote the manuscript. The author supervised the master thesis that was the basis for this publication.

VI As the main and leading author of this publication, the author conceived the idea of measuring cyberethical behaviour using virtual labs. The author contributed to the design and conduct of the experiment, and analysis of the results. The author also prepared the figures, and wrote the manuscript. The author supervised the master thesis that was the basis for this publication.

VII As the main and leading author of this publication, the author designed and conducted the introductory cybersecurity course that was the basis for this study. The author collected the data, analysed the results, prepared the figures, and wrote the manuscript.

Other publications

VIII K. Maennel, S. Mäses, and O. Maennel. Cyber hygiene: The big picture. In *Nordic Conference on Secure IT Systems*, pages 218–226. Springer, 2018

IX K. Maennel, S. Mäses, S. Sütterlin, M. Ernits, and O. Maennel. Using technical cybersecurity exercises in university admissions and skill evaluation. *IFAC-PapersOnLine*, 52(19):169–174, 2019

Set your life on fire.
Seek those who fan your flames.
—Rumi

Build a man a fire, and he’ll be warm for a day.
Set a man on fire, and he’ll be warm for the rest of his life.
—Terry Pratchett

Victory comes from finding opportunities in problems.
—Sun Tzu

1 Introduction

This thesis demonstrates how to improve cybersecurity education with the use of virtual simulation exercises that evaluate specific cybersecurity-related competences. As interconnected computers are progressively involved in various everyday processes [76], it is essential to have people with adequate competencies to operate various machines in an efficient and secure way. In addition to the high demand for cybersecurity specialists [32], every employee dealing with computers is increasingly expected to be familiar with relevant cybersecurity practices such as choosing strong passwords and reporting phishing emails [54].

The rising complexity of computer networks has made it increasingly difficult to ensure that the systems are used securely. Furthermore, the production of new machines and software is much faster than training skilled people to securely manage those devices. That has led to a serious shortage of cybersecurity competencies [32]. For overcoming this lack of cybersecurity competencies, cybersecurity education must become more effective.

The effectiveness of training can be improved by utilising learning theories that provide verified instructional strategies and techniques for facilitating effective learning. Various systematic approaches to understanding and describing learning such as behaviourism, cognitivism, and constructivism overlap in several aspects [29]. For example, they all agree on the importance of feedback. Behaviourists use feedback for behavioural reinforcement, cognitivists for supporting accurate mental connections, and constructivists for evaluating the understandings constructed by the learners. Similarly, it is generally agreed that structured guidance and meaningful hands-on tasks facilitate effective learning process [29].

Although there are some generally accepted approaches for designing effective education, the success of a particular approach is nevertheless heavily context dependent [8]. **To ensure the effectiveness of education in a particular context, learners must be evaluated regularly.** There are factors, though, that make it challenging to measure how skilled someone is in cybersecurity.

The cybersecurity field is still relatively young and therefore not clearly defined [85]. In most countries there are no dedicated cybersecurity lessons given in primary and secondary education. Also, in universities, cybersecurity concepts are often taught as a part of other courses such as computer networking or programming. Therefore, looking at the transcripts or grades of a student is not likely to give an accurate understanding about their cybersecurity-related skills.

Knowledge tests that are frequently used in education do not predict the proficiency in the actual cybersecurity skills relevant to job performance [48]. For example, someone might not remember the exact command to finish a task but is able to find this command from the Internet or by using in-built help functionality in less than a minute. Knowledge acquisition skill (e.g., to use an efficient search strategy) that is highly valued in real-life situations, is commonly not valued or even frowned upon while taking a knowledge test. Therefore, there is a mismatch between the competence that is tested (retrieving factual knowledge) and the competence that is actually expected (finding knowledge, applying skills) in the job market.

Work sample testing enables practical skills to be evaluated by giving people practical tasks to solve [86]. Practical work tasks integrate the application of theoretical knowledge, skills of information search and processing by adapting existing knowledge to novel problem solving. Thus, a realistic simulation of a work sample can arguably provide a high level of predictive and external validity. Using practical tasks also helps to avoid socially desired responding, that is common in questionnaires and interviews [49].

Cybersecurity context is especially suitable for using simulations. Due to the digital nature of cybersecurity, the experience gained from a virtual simulation can relatively easily be transferred to a similar task in the real-life work environment [77]. Furthermore, the recent developments in virtualisation technologies have allowed to fasten the creation and management of custom computer networks. Not only it is possible to create virtual networks on-demand, but also to have automated scoring scripts evaluate the individual performance. This has enabled scalable hands-on virtual labs to be created where participants can demonstrate their cybersecurity skills [27, 50, 95].

These virtual labs could be seen as a type of serious games, because they combine participatory experience of simulations with motivational aspects of gaming and gamification, such as task engagement and a high degree of immersion [53]. Serious cybersecurity games seem to be a good way for engaging the participants whilst at the same time training and measuring their cybersecurity-related skills.

However, the scoring system of those serious cybersecurity games is often quite generic and lacking any connection to specific work role or skill set. Thus, it is not possible to analyse which participant would be a better fit for a specific team or job role. Without proper measurement of cybersecurity-related skills, it is not possible to analyse the impact of the serious games nor other types of cybersecurity education. Organisations invest millions educating employees to be skilled in cybersecurity. Educators also spend significant time and effort finding the most effective ways to teach people about cybersecurity. However, to optimise those efforts and understand the actual impact, regular cybersecurity skill evaluations are needed.

This thesis provides guidance for creating more meaningful cybersecurity exercises. Different ways are explored for creating serious cybersecurity simulations that measure competencies and provide more meaningful feedback than just stating the final amount of gathered points. For example, it is shown how skill evaluations in virtual simulations can be mapped to specific job roles defined by NIST NICE framework [66]. Established individual competency profiles can contribute to early detection and intervention by targeting insufficient competency areas and identify or build upon existing strength to facilitate specialisation (Publication VII). The competency-driven exercise design approach suggested in this thesis is put in practice by several specifically created virtual simulations. In addition to illustrating the feasibility of competency-driven exercise design, those virtual simulations introduce novel technical solutions that demonstrate the versatility of serious games as an educational medium. The structured way of automatically evaluating competencies in virtual simulations provides a scalable way to build up a more effective, competency-based cybersecurity education.

1.1 Key terms

This section discusses how various key terms are used in the current context.

Evaluation is commonly defined as the determination of the level of the quality [6] with the aim of providing a judgement [43]. Related terms include assessment, measurement, test—many of which are used interchangeably in the common language. The term **measurement** signifies assigning numerical value to some observed characteristic. In contrast to evaluation, **assessment** tends to have a more diagnostic nature and is focusing on improving the performance [6], not providing judgement. Test is a form of assessment. Evaluation and assessment are often used together in an educational setting. A good teacher gives students both a grade (evaluation) and feedback how to improve (assessment). This thesis looks at different for gathering and interpreting measurements systematically. Therefore, the terms evaluation, assessment, and test are often used to emphasise a specific viewpoint while looking at the same process.

The term **cybersecurity** consists of the prefix “cyber-” and the term “security”—both of which reveal a great amount of complexity and ambiguity in a closer look. While in common language, cybersecurity is used interchangeably with computer security and information security, all those terms can mean different things to various specialists. This thesis uses the term cybersecurity as proposed by Schatz et al. after conducting lexical and semantic analysis on various definitions: *“The approach and actions associated with security risk management processes followed by organizations and states to protect confidentiality, integrity and availability of data and assets used in cyber space. The concept includes guidelines, policies and collections of safeguards, technologies, tools and training to provide the best protection for the state of the cyber environment and its users.”* [83]

The use of the terms **competency** and competence can be confusingly ambiguous [30, 92]. While some authors distinguish the terms competence and competency, they are used interchangeably in the current context. Most of the literature agrees that competency consists of knowledge, skills, and “something else”. Regarding that “something else” part, there are various suggestions, e.g., ability, aptitude, attitude, attribute, and/or dispositions [92]. The current paper is aligned with research considering the **competency as a set of knowledge, skills, and dispositions** [30, 72, 82]. More specifically, the main focus is put on measurements of skills relevant to cybersecurity.

Various cybersecurity situations can be looked at as a game where various participants (players) are competing to achieve their goals (winning condition) in a set of constraints (rules) [51]. The term **serious games** [18] is often used to signify games that are designed with the main aim being other than entertainment—usually learning and behaviour change [23]. It should be noted that whether the seriousness of the game is defined by the (rarely accessible) intention of its developer(s) or whether serious game is defined as any game used for more than entertainment—every game may be seen as a serious game [44]. In this thesis, the term “serious games” signifies the use of gamified simulations for educational purposes (in the context of measuring skills and behaviour). Educational use of video games is not in the scope of this thesis.

Gamification means using game-like elements in non-game situations, mostly with the aim of increasing the engagement and motivation of the participants [36]. According to the wider definition of gamification, it can be understood as the concept of human-centered design in general [21].

Simulation can be defined as a “*representation of the function, operation or features of one process or system through the use of another*” [51]. Therefore, simulations imitate the behaviour of some situation without the actual consequences. For example, a flight simulator enables the pilot to test herself without the fear and cost of actually crashing an airplane. Similarly, virtual cybersecurity labs enable participants to test their cybersecurity skills in a safe sandbox environment [27]. An **exercise** is understood in this context as a set of activities designed to train, develop, or test one’s competency in a particular field. Cybersecurity-related simulation exercises often take place using a virtual lab. The term **virtual lab** signifies a set of virtual machines that are remotely accessible and used for a simulation. When the simulation is imitating the real work tasks realistically enough, then it is similar to **work sample tests**. Work sample tests are believed to be among the most valid predictors of job performance [78].

To sum up, this thesis shows how to combine elements from work sample tests, serious games, and gamification—enabling the creation of simulations of various cybersecurity tasks, that are engaging and effective at measuring related competencies.

1.2 Problem statement

There is the growing need for a workforce with cybersecurity competencies [93]. Lack of proper competence can lead to various unwanted outcomes such as leaks of sensitive data, and dangerous vulnerabilities in medical machinery and industrial control devices. At the same time, educational organisations struggle with providing relevant, scalable, hands-on skills that would take into account personal preferences and differences [32]. Cybersecurity exercises (simulations, labs, serious games) seem to be a very appropriate way for teaching and evaluating relevant skills. However, the full potential of cybersecurity simulations seems to be underutilised.

As a metaphor[1], virtual simulation exercises could be compared to cars. Used correctly, they can take us quickly and efficiently to various destinations. At the moment though, most simulation exercises are treated more like chariots. Imagine a golden Porsche pulled by two horses (Figure 1). It might be more comfortable than a traditional chariot, but not utilizing the full potential of the Porsche. Similarly, simulation exercises could be used for much more in various educational contexts. This thesis explores various ways how to use virtual simulation exercises to evaluate cybersecurity competences.

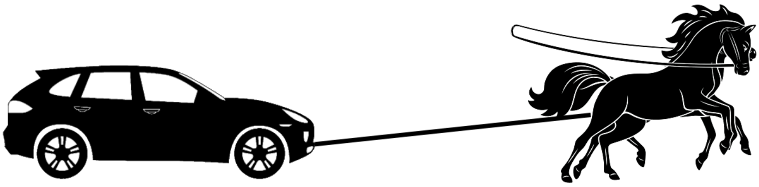


Figure 1 – Horses pulling a car like a chariot—used here as a metaphor illustrating the underutilised potential of virtual simulation exercises.

Although various cybersecurity exercises give participants hands-on experience with practical tasks, the results are rarely measured in a constructive way. There are classroom exercises that contribute to the final grade, and there are extracurricular competitions that focus on the comparative ranking. Those exercises could be more insightful when developed in a more competency-oriented way.

Competency-based learning (CBL) and competency-based education are increasingly gaining traction during the last decade. CBL has proven to be effective and engaging [38] for developing specific competencies. However, before it is possible to focus on specific competencies, those competencies need to be identified. In cybersecurity, there are several initiatives such as NIST NICE [66], CyBOK [35], and others [32] that aim to identify and classify relevant competencies. While those frameworks are often used as curricular guidance (especially CAE [65] and CSEC [91]) they could also be used for structuring individual skill evaluations.

In a classroom setting, a more detailed competency-based feedback would enable students to better understand the strengths needed in their potential future job. Now, getting a medium grade for a course does not indicate what learning outcomes were achieved and how are those learning outcomes connected to potential job roles. A competency-based learning would enable students to set specific individual learning goals and get feedback regarding their progress towards those goals.

Evaluating cybersecurity skills is essential for identifying and educating the workforce for organisations that are battling increasing cybersecurity threats. The aim of this thesis

[1] Analogies and metaphors are examples of an effective cognitive strategy of connecting new information with existing knowledge in some meaningful way [29].

is to find out how to use virtual simulation exercises for evaluating cybersecurity-related competences with the focus on cybersecurity-related skills. That gives the basis for evaluating cybersecurity-related skills in a practical and scalable way.

1.3 Research questions

This thesis focuses on evaluating cybersecurity-related competences (especially skills and behaviour) using simulation exercises. An effective cybersecurity exercise incorporates practical virtual simulations that are aligned with cybersecurity frameworks and are evaluated according to an established educational taxonomy. To find out how to evaluate cybersecurity-related competences through simulation exercises, the following research questions are defined.

- How to categorise cybersecurity-related competences? (RQ 1)
- What kind of metrics relevant to skills evaluation can be gathered using simulation exercises? (RQ 2)
- What kind of cybersecurity-related skills can be evaluated with simulation exercises? (RQ 3)
- How to design simulation exercises suitable for measuring cybersecurity-related skills? (RQ 4)

Together, the answers to the research questions explain how to create an effective cybersecurity simulation exercise. In section 1.4, the connections between research questions and publications are given.

1.4 Contribution

There are numerous cybersecurity competitions and technical exercises, but they usually do not evaluate the specific skills of the participants. One participant of a cybersecurity competition might get more points than another entrant, but it is rarely possible to do a further skill comparison between those competitors. This thesis shows how to design and implement virtual simulations that enable the skills of participants to be evaluated in a way that is scalable and comparable to the results of other similarly designed exercises. This thesis is based on 7 publications that can be grouped into three sets (as illustrated by the Figure 2):

- Publications I, II, and III form the foundational structure of skill measurement and competency-driven exercise design. First, the argumentation behind skill measurements using virtual simulations is given (Publications I and II) and a Cybersec-Tech window is suggested for a categorisation of cybersecurity-related competencies (Publication II). A comprehensive structured approach for creating competency-driven cybersecurity exercises is then suggested (Publications II and III). The research questions addressed in this section are RQ 1, RQ 2, and RQ 4.
- Publications IV, V, and VI present case studies of specific simulations that measure skills from different sections of the Cybersec-Tech window. First, non-technical and cybersecurity-related cybersecurity awareness is measured with phishing simulations (Publication IV). Second, a novel way of simulating realistic OSINT (Open Source Intelligence) tasks in the exercise environment is presented (Publication V).

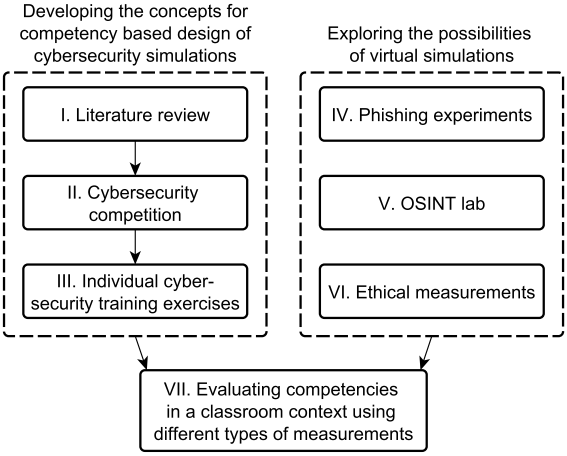

Developing the concepts for competency based design of cybersecurity simulations
Exploring the possibilities of virtual simulations
I. Literature review
II. Cybersecurity competition
III. Individual cyber-security training exercises
IV. Phishing experiments
V. OSINT lab
VI. Ethical measurements
VII. Evaluating competencies in a classroom context using different types of measurements

Figure 2 – Illustrating the context connecting the publications (I..VII) that are the basis for this thesis

Finally, it is shown how virtual simulations can be used for measuring ethical behaviour, which is essentially non-technical and not directly cybersecurity-related, in a cybersecurity context (Publication VI). The research questions addressed in this section are RQ 2 and RQ 3.

- Publication VII presents a case study of a holistic approach—using cybersecurity exercises in a classroom setting where metrics from virtual simulations are combined with other measurements (Publication VII). This section covers all the research questions (RQ 1, RQ 2, RQ 3, RQ 4).

Altogether, this thesis claims that virtual simulation exercises (serious games) are a seminal medium for cybersecurity education when used in a structured manner. General competency-based exercise design processes are formalised (Chapter 4), providing guidance for a more effective cybersecurity education. Novel virtual simulations are constructed and analysed (Chapter 5). In conclusion, a case study is presented (Chapter 6) where metrics from virtual simulations are used together with other types of metrics in the context of a bachelor (undergraduate) level introductory cybersecurity course.

2 Related work

An effective hands-on cybersecurity simulation should reach beyond purely the technical scope of its implementation. It should also be connected to a structured body of competencies and have meaningful scales of evaluation (as illustrated by the area marked by the eye symbol in Figure 3).

Currently, competency frameworks, virtual simulations, and proficiency scales are all occasionally used in cybersecurity education, but not in a unified manner (as illustrated in the top part of Figure 3). Competency frameworks are used for general curricular guidance but not for individual skill assessment. Virtual simulations are used for practical assignments but rarely graded on a scale of a meaningful educational taxonomy. While general course outcomes might be described following educational taxonomies, the individual results of virtual simulations are usually not.

The next sections look deeper into various cybersecurity frameworks, exercises and educational taxonomies.

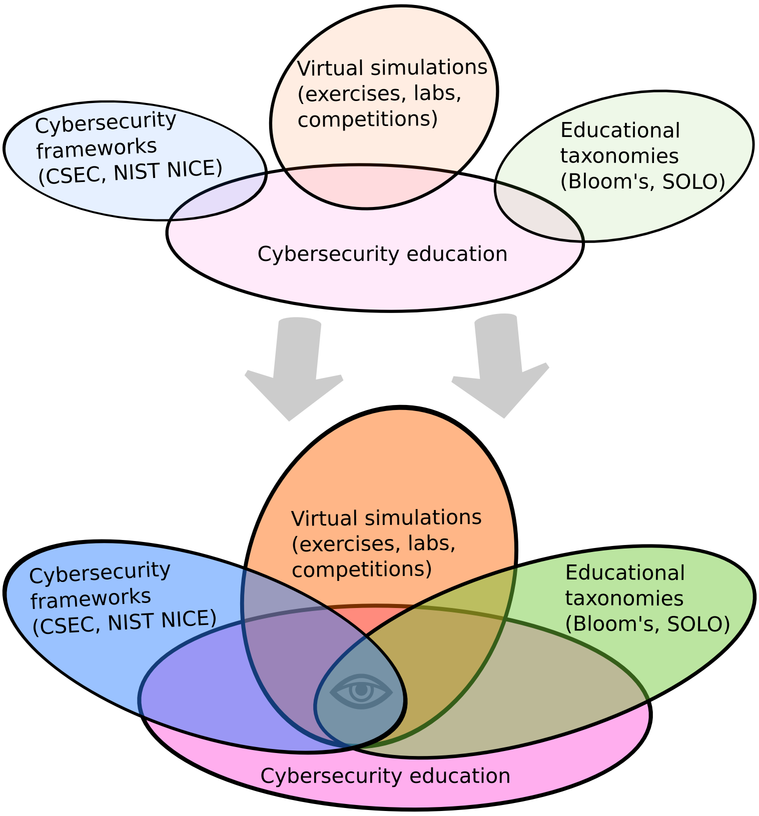

Virtual simulations (exercises, labs, competitions)
Cybersecurity frameworks (CSEC, NIST NICE)
Educational taxonomies (Bloom's, SOLO)
Cybersecurity education
Virtual simulations (exercises, labs, competitions)
Cybersecurity frameworks (CSEC, NIST NICE)
Educational taxonomies (Bloom's, SOLO)
Cybersecurity education

Figure 3 – The current state (top) and envisioned state with wider interdisciplinary synergies (bottom). An effective cybersecurity exercise is incorporating practical virtual simulations that are aligned with cybersecurity frameworks and evaluated according to an established educational taxonomy. The intersection of those areas is marked by the eye symbol. Best viewed in colour.

2.1 Cybersecurity education and frameworks

Cybersecurity is a relatively new field and most curricula specializing on it were established after 2012 [17]. Due to the term "cybersecurity" being quite wide and ambiguous, different academic programmes can cover dramatically different content—even if the resulting degree has the exact same name [32]. In addition to academic programmes, there are over 100 industry providers for various cybersecurity qualifications—some more general and others focusing on specific technologies [32].

To have a more unified approach to the field of cybersecurity, there have been multiple initiatives aiming to identify and classify relevant competencies and topic areas. NIST has published National Initiative for Cybersecurity Education Cybersecurity Workforce Framework (NICE framework) [66] that defines the knowledge, skills, and abilities (KSA) needed for cybersecurity-related jobs. It also defines tasks connected to job roles. The National Security Agency (NSA) of US has created Centers of Academic Excellence (CAE) [65] providing a list of knowledge units that are used in designated specialisations of cyber defence and cyber operations. A joint task force led by ACM has developed curricular guidelines for cybersecurity curricula (CSEC) [91] that also provides the connecting mappings with CAE and NICE frameworks.

The NICE framework tends to be the most widely used both in research and in the private sector. Several studies have looked into finding the most important KSA from NICE framework. Amstrong et al. [4] have investigated relevant KSA for jobs in vulnerability assessment and management. They have brought out 12 out of the selected 31 KSA of NIST NICE [66] that were rated to be significantly more important than neutral to participants' jobs. Jones et al. [45] identified the most important KSA areas from NIST NICE framework based on expert interviews. 15 out of the selected 32 KSA were found to be *"significantly more important than neutral after a series of one-sample t-tests"* [45]. In addition to technical skills, they stress the equal importance of non-technical skills such as communication and collaboration. NICE framework also lists relevant non-technical skills such as *S0344—Skill to prepare and deliver reports*, *presentations and briefings*, *to include using visual aids or presentation technology* and *S0359—Skill to use critical thinking to analyze organizational patterns and relationships* [66].

While the NICE framework focuses on KSA relevant to cybersecurity specialists, every computer user needs less technical skills such as choosing a strong password and recognising a malicious e-mail for staying secure in their everyday life. The categorisation of less technical cybersecurity-related skills (e.g., [19]) is aligned with seven focus areas mentioned by Parsons et al. [73] that include password management, email use, Internet use, social media use, mobile devices, information handling, and incident reporting. Additionally, there are focus areas of updating software and backing up data, that Parsons et al. [73] consider to be the responsibility of the IT personnel, not of the regular user. Often terms such as security awareness [37], cybersecurity culture [33], and cyber hygiene [54] are used in connection with the less technical cybersecurity-related skills, although there are no agreed definitions of those terms.

In general, it could be said that the combined efforts of initiatives such as CAE, CSEC and NICE have produced a usable high-level structure for relevant cybersecurity skills. At the same time those systematic approaches are not used much in educational programmes. While the cybersecurity frameworks are used for high-level guidance for including specific topics into the curricula, those frameworks are rarely used to provide a more detailed view into personal study outcomes. This thesis argues that these type of well-established frameworks could also provide the structure for a more detailed skill evaluation, for example in the context of cybersecurity exercises.

2.2 Cybersecurity exercises

The idea of creating technical cybersecurity exercises is not new. There were some ideas about cybersecurity exercises in the end of 1990s, although mostly in the form of laboratory projects [52]. In the first half of 2000s, the concept of capture the flag (CTF) was taken from the military simulation of real-world battle situations [5] to cyberspace [24, 25].

The basic idea of a CTF game is to have two teams protecting their physical flags that are fixed in one location while at the same time reaching the flags of the opposing team(s) [52]. In cyberspace, the flags are often computer systems, services or pieces of information that one team has to protect, and others are trying to get access to. The CTF format enables a great amount of flexibility for the organisers designing the exercise and for the participants in finding ways to achieve the given objective [25].

Starting from the beginning of 2000s, cybersecurity exercises have grown in both variety and complexity. There are both individual exercises and team exercises (with usually the Red team as attackers and Blue teams as defenders of the systems). While the main purpose of the exercises is usually to give participants the opportunity to demonstrate and gain hands-on cybersecurity skills, the exercises have also been found as a great resource for gathering data for research. Cybersecurity exercises have been used for empirical validation of the frameworks, methods, tools, and principles found in literature [88].

The current thesis aligns well with recent research in cybersecurity exercises that is increasingly aimed at better understanding the educational and psychological side. As cybersecurity exercises are increasingly used in different educational settings [96, 63], there is a growing interest in connecting those exercises more with educational learning outcomes [20] and understanding other learning aspects (such as learner's motivation [1]). Existing research stresses the importance of putting proper systems and procedures in place to effectively monitor and quantify team behaviour for valid analyses [39]. Increasing amounts of measurements are collected during the exercises. For example, timestamps have been analysed to assess team performance [55], and command line history to qualitatively assess and compare team performance [94].

Nevertheless, the gathering of various metrics from simulations is still in its early stage of development and there is much room for improvement. This thesis investigates how to better use cybersecurity exercises to evaluate cybersecurity-related competencies.

2.3 Educational taxonomies

The metrics gathered during cybersecurity simulations could benefit from following a structure of an established educational taxonomy. Taxonomies of educational objectives can be used for communicating accurately the learning outcomes and performance in assessments [31].

For classroom education, the most widely used one tends to be Bloom's taxonomy [12] or its revised version [2]. The original Bloom's taxonomy of the cognitive domain lists six sequential categories following the learning process: knowledge, comprehension, application, analysis, synthesis, evaluation.

The revised Bloom's taxonomy of the cognitive domain changed the order of the two highest categories and assigned verbs as category names: remember, understand, apply, analyse, evaluate, and create. The revised Bloom's taxonomy also defined a separate knowledge dimension listing four different types of knowledge: factual, conceptual, procedural, and metacognitive [31]. There are several taxonomies that build upon Bloom's ideas. For example, Niemierko's ABC taxonomy unites the three higher level Bloom's categories and differentiates knowledge from abilities and skills [67]. Also, a matrix taxonomy has been suggested that separates Bloom's six levels into two dimensions: producing (incorporating apply and create), and interpreting (incorporating remember, understand, analyse and evaluate) [31].

Another widely known educational taxonomy is the Structure of the Observed Learning Outcome (SOLO) [10]. The SOLO levels classify learning outcomes based on their complexity: prestructural, unistructural, multistructural, relational, extended abstract. While SOLO has been used for analysing progression of science curricula in some universities [13], the prominent taxonomy used for computer science is the revised Bloom's taxonomy [31]. In addition to its implementation in computer science courses [89], it has also been embedded into CSEC2020 curriculum [91] providing guidance for cybersecurity curricula.

In conclusion, there are various educational taxonomies that could be used for evaluating the virtual simulation tasks and connected proficiencies in a more comparable and meaningful way. Later sections explore a competency-driven design process that uses educational taxonomies for establishing meaningful proficiency levels.

3 Methodology

The main aim of this thesis is to improve the usage of virtual simulation exercises in a cybersecurity skill evaluation context. To achieve this aim, this thesis explores ways how to enhance practical virtual simulations by aligning them with cybersecurity frameworks and established educational taxonomies. As previous research does not provide a sufficient basis for defining specific quantitative hypotheses, a more foundational approach is needed.

Design science deals with "innovations that define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, and use of information systems can be effectively and efficiently accomplished" [40] and is therefore suitable to address the research questions of this thesis. This thesis looks at multiple cases where some sort of method is being designed together with a proof of concept software solution. The design science research methodology (DSRM) process by Peffers et al. [75] is followed to present these cases in a structured manner. While the described cases provide novel insights to design processes of cybersecurity simulation exercises, the causal analysis and statistically generalisable results are out of the scope of this thesis.

It should be noted that many software developing [81] and other design-oriented approaches in the information systems area tend to be generally similar (as illustrated by Figure 4). As developing a system often aims to solve some sort of problem, it is not surprising to find that several software developing approaches and models are well aligned with the general process of problem solving (Figure 4).

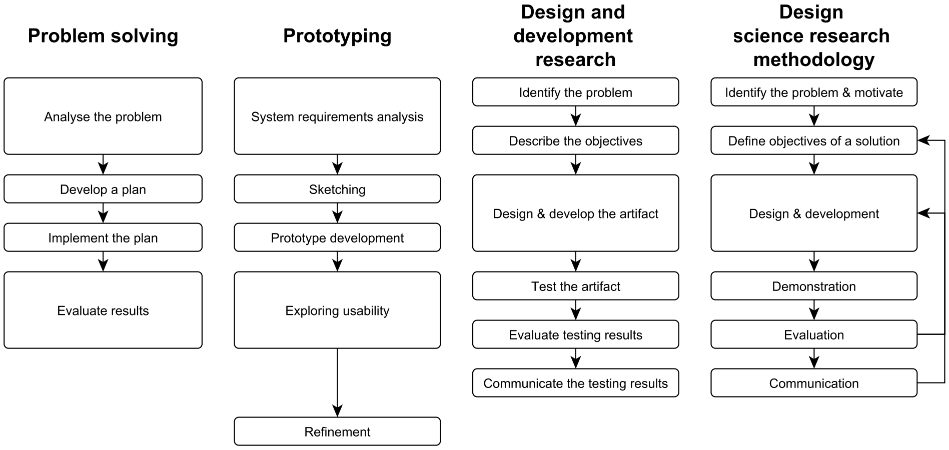


Figure 4 – Comparing the generic problem solving process [64] to various approaches used for developing systems: prototyping [64], design and development research [26], and design science research methodology [75]

The development of proof-of-concept systems can also be seen as prototyping. Prototyping involves the creation of a simplified model or sample to test a concept or a process [7]. Using an expert system prototypes can be analogous to a case study [69]. Both are used to obtain more insights into a product or process. Prototyping can also be viewed as an underlying approach for design science research where implementing and evaluating an information system is an element of the research activity [7].

Design science research itself is generally considered to be a paradigm instead of a discrete research methodology [7, 41]. Nevertheless, there is strong support for design-oriented information systems research [75, 26, 70]. Overall, this dissertation follows the DSRM process [75] that consists of the following phases:

1. identify problem & motivate,
2. define objectives of a solution,
3. design & development,
4. demonstration,
5. evaluation,
6. communication.

The first 5 phases are brought out in the following chapters of the thesis that are dealing with the specific cases where various artefacts were developed. The last phase in DSRM—communication—is carried out by this thesis in more general terms, and by the publications that it is based on. Figure 5 illustrates the application of the DSRM in this dissertation.

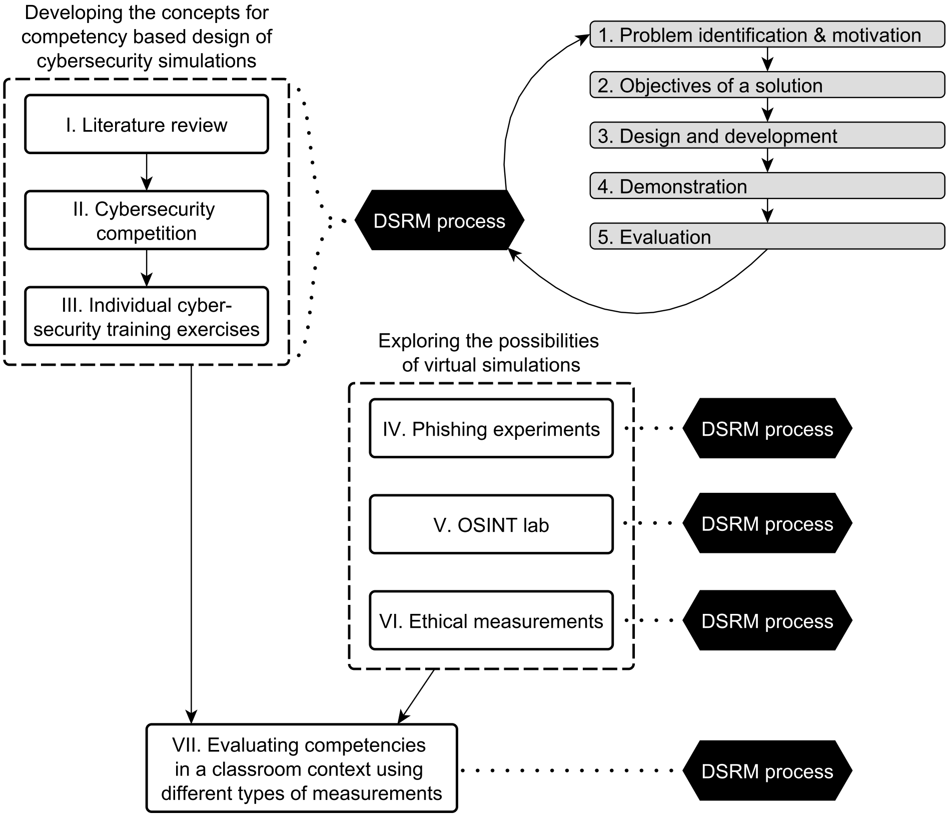

Developing the concepts for competency based design of cybersecurity simulations
I. Literature review
II. Cybersecurity competition
III. Individual cyber-security training exercises
DSRM process
1. Problem identification & motivation
2. Objectives of a solution
3. Design and development
4. Demonstration
5. Evaluation
Exploring the possibilities of virtual simulations
IV. Phishing experiments
DSRM process
V. OSINT lab
DSRM process
VI. Ethical measurements
DSRM process
VII. Evaluating competencies in a classroom context using different types of measurements
DSRM process

Figure 5 – Illustrating the application of the design science research methodology (DSRM) process [75] throughout the different cases, each DSRM consisting of 5 action phases. Roman numerals mark the underlying publications and Latin numbers signify the phases in the DSRM process.

4 Designing competency-based cybersecurity simulations

4.1 Problem identification and motivation

There is a need for effective and efficient cybersecurity education that is scalable, personal, and hands-on. Virtual cybersecurity simulation exercises are suitable for taking cybersecurity education to a new level. However, the use of simulation exercises is still new and relatively little researched. While it is clear that cybersecurity exercises enable the measurement of various skill-related metrics, it is still rarely done in practice. Therefore, the topic of cybersecurity exercises needs exploratory research to describe and understand the current state and generate new ideas for new research [80].

A more structured approach for designing and implementing competency-based cybersecurity exercises would help to decrease the growing gap in cybersecurity skills [32]. If participants' progress can be measured, then it would be possible to find optimal learning paths and best suited teaching methodologies. Therefore, evaluating competencies is an essential prerequisite for effective education. Furthermore, an accurate evaluation of competencies can also help to locate suitable workforce to cybersecurity-related job positions.

4.2 Objectives of a solution

The objective was to create a method for designing cybersecurity exercises to effectively measure relevant competencies. The design and implementation concepts should be general enough to be applicable for competitions as well as for individual training. The following requirements were defined:

- describe the cybersecurity competency-based exercise life cycle process,
- identify specific design approaches for connecting exercise measurements with competency evaluations.

4.3 Design and development

Desk research was used to analyse the life cycle process of existing cybersecurity exercises and a list of relevant metrics was compiled (Publication I). Various suggested processes for conducting a cybersecurity exercise tend to follow a linear approach. The processes mentioned in the literature are described as a sequence of non-iterative phases with little attention to content design. The process by ENISA [71] only gives a very high-level overview from the project manager's perspective. Patriciu et al. [74] mention the importance of high-level objectives (offensive security and/or defensive security), and that the technical system design should support the exercise objectives. Wilhemson et al. [97] mention that exercise participants can provide their ideas about knowledge and skills in the exercise design phase. However, no skill-specific evaluation is discussed in detail.

These results from desk research were rather surprising because modern system development processes usually follow an iterative flow [81]. Discussions with people having experience in organising international cybersecurity exercises revealed that the actual design process of an exercise is much more iterative in practice. After analysing the process models in literature and inputs from the practitioners, a new 10-step model was created (Figure 6). This model takes into account the philosophy of agile software development [81] and focuses on the two main components of the exercise: technical design and competency model. The model is described more in detail in Publication II.

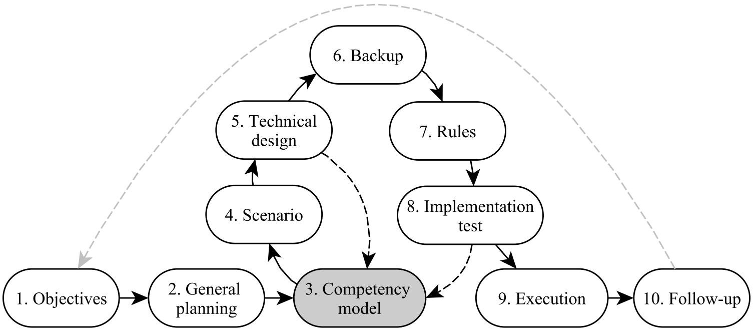


Figure 6 – Suggested 10-step process for designing and implementing cybersecurity exercises

While technical design is an implicit part of every technical cybersecurity exercise, little attention is put to connecting results with specific competencies. Most exercises track the progress of participants with points gained for completing the tasks. However, it is usually impossible to infer any exact information about individual competencies from this generic score. To generate a more specific competency profile for each participant, the specific measurement points should be categorised and connected to relevant skills. Figure 7 shows an example illustrating how the result of an attack can be an indication for the system administration competency. Instead of simply adding up the scores gathered from measurement points (or measurement categories) to see who scores higher, this kind of exercise design enables deeper insights about the related competencies to be gained.

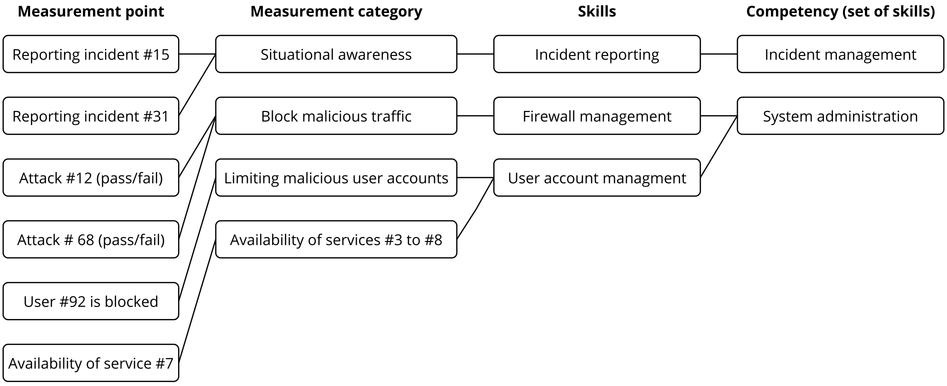


Figure 7 – Example illustrating how measurement points (on the left) can be connected to specific competencies (on the right)

Depending on the context, a failed attack might not always correctly indicate a high-level system administration competency. If the system administrator switches off a service completely, then attacks against this service will fail. Nevertheless, it does not reflect the competency of defending against such attacks. To adequately evaluate a competency, multiple events often need to be looked at together. Publication III [28] introduces the concept of a game event log (GEL) to capture various events happening during the serious game.

As an example, to prove a competency of defending a specific system against an attack, several events must happen in a particular pattern. First, an availability check must succeed showing that the targeted system is functional and available. Then, an attack against this system must fail. Finally, another availability check must show that the system is still functional after the failed attack. If any of those events has a different outcome, then the defending competency cannot be inferred.

For different event patterns to reflect specific competencies, a consistent mapping must be done throughout the exercise (Figure 8). The mapping process should follow the conceptual design of the exercise, specifying the competencies and skills to be evaluated. Afterwards, specific tasks can be identified for evaluating the defined skills. The technical design is then used for defining relevant game events and creating suitable exercise environment. If this design process (Figure 8) is properly followed, relevant actions in the simulation automatically infer specific competencies. The competency driven design process, including the usage of GEL, is described more in detail in Publication III.

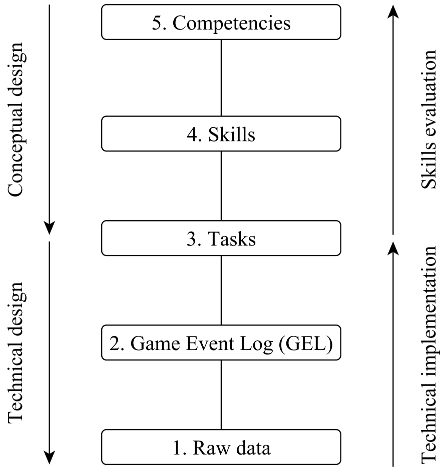


Figure 8 – Competency driven design

4.4 Demonstration

Two case studies were conducted to specify and test the suggested processes. One case involved a national level cybersecurity competition for students. The other case explored the exercise design in the context of commercial cybersecurity trainings.

The actual design and implementation process of the national level cybersecurity competition (Publication II) was well aligned with the suggested 10-step model (Figure 6). Predefining relevant competency areas helped to form a relevant scenario and technical design. Some topics, such as attacking simulated Internet of things devices, were excluded from scope due to high implementation cost. It took several iterations between competency model, scenario, and technical design to finalise the exercise setup. The implementation of revised Bloom's taxonomy [2] resulted in most tasks categorised as applying procedural knowledge and a few connected to analysis level. The NIST NICE framework [66] was used for high-level guidance. Work roles were analysed in the context of the exercise, but more specific lists of skills and abilities were disregarded.

In the other case study, 27 participants went through 13 different virtual labs on the topic of web application security (Publication III). Additionally, another virtual lab was used for pre-test and post-test to evaluate the learning progress. Using the suggested competency-driven design process (Figure 8), every participant received automated feedback on their specific skills. Although implementing various game event logging scripts and connecting them to specific skills took time in the beginning, it helped to create a system that can now provide skill-specific feedback in a scalable way.

4.5 Evaluation

The objective to describe the cybersecurity exercise life cycle process was achieved. The literature review revealed the lack of iterative approaches and little attention paid to measuring specific competencies. Therefore, a case study was conducted to identify and specify the phases needed for cybersecurity exercise design and implementation. As a result, an iterative 10-step process model (Figure 6) was created that also includes competency modelling as a separate phase. For specifying the steps needed to connect exercise measurements with competency evaluations, the second case study was carried out. The competency-driven exercise design process was formalised as shown in Figure 8.

Overall, it was shown that virtual simulations can be used for teaching and evaluating both simple and complex skills with various proficiency levels. Short and targeted mini-exercises are good for teaching and evaluating specific skills, such as the use of a particular software tool. Longer and more complex exercises enable to evaluate higher level of skills such as adapting the existing knowledge to a new context. The complex exercises often have a general goal that can be achieved in various ways. While it allows participants to demonstrate higher level competencies, it also makes it more difficult to understand what skill was used. For example, a malicious attack towards a web site could be blocked by configuring a firewall or patching the vulnerability in the source code of the web site. Both approaches can be successful, but the underlying competencies could differ significantly. Furthermore, both approaches might be completed semi-randomly using a trial and error approach. While guessing a suitable solution might work in particular context, it does not mean that the participant possesses required analytical competencies to be successful in a similar context. Therefore, more complex exercises require a closer monitoring of participants' activities for evaluating their specific skills.

In conclusion, the suggested process was considered useful as it enabled fast feedback to be given regarding individual skills in a scalable way.

5 Simulation cases

Cybersecurity simulations can be used to teach how to attack a vulnerable web server, but they can also be used for much more. This section explores the possibilities related to the use of different simulations. First, a model is introduced for categorising different skills needed in cybersecurity. Then, three separate cases are described that demonstrate the versatility of simulations.

5.1 Cybersec-Tech window

To evaluate cybersecurity-related competences, those must first be defined and classified (RQ 1). For analysing the theme of cybersecurity-related competences, a content analysis approach was used to operationalise it (Publication II). As a result, the conceptual map of cybersecurity-related competencies was formed (Figure 9). The Cybersec-Tech window is suggested as a generic model to distinguish different skills needed in cybersecurity field.

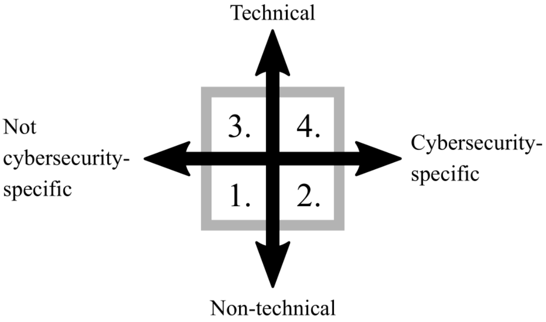


Figure 9 – Cybersec-Tech window for categorising different skills needed in cybersecurity field

The Cybersec-Tech window shown in Figure 9 illustrates how the different cybersecurity-related competences (including knowledge, and skills) can be divided into four categories (Publication II).

1. Skills that are non-technical and not cybersecurity-specific. This includes universally transferable skills such as communication and leadership. While in our context, those skills are not considered technical, nor cybersecurity-specific, they are still highly valued. Tabletop exercises are often created to address this quadrant of Cybersec-Tech window. Exercises that are carried out in teams, require teamwork skills that are also included in this quadrant.

2. Second quadrant includes skills that are cybersecurity-specific, but non-technical. Security awareness trainings often target those skills. Not opening suspicious links in phishing emails or creating a secure password might not require very technical skills. Nevertheless, it is important to distinguish phishing emails and behaving according to the security policy for ensuring a good security posture.

3. Third quadrant consists of skills that are technical, but not necessarily cybersecurity-related. Coding in some specific programming language is a good example of a skill belonging to this area.

4. Fourth quadrant consists of skills that are both technical and cybersecurity-specific. Implementing different encryption algorithms or performing a SQL injection attack are some samples of those type of skills.

Technical cybersecurity exercises target mostly 3rd and 4th quadrant. Frameworks focusing on cybersecurity-related job descriptions (such as NIST NICE [66]) also focus on those quadrants and describe mostly technical competences. However, there are several non-technical competencies that are also very relevant to cybersecurity such as skills to communicate effectively or behave ethically.

The following sections describe three different simulations that are measuring skills from different Cybersec-Tech window quadrants. As the topic of this thesis is focusing on cybersecurity-related competencies, then it is only natural to cover quadrants 2 and 4 that are targeting cybersecurity-specific skills. Additionally, a method of measuring cyberethical behaviour is described (quadrant 1) to illustrate how various non-technical measurements can be added to virtual labs.

5.2 Phishing simulation (quadrant 2)

5.2.1 Problem identification and motivation

The non-technical skills connected to cybersecurity are mostly generic transferable skills (such as leadership or teamworking skills) or those connected to cybersecurity awareness and cyber hygiene. The second quadrant of the Cybersec-Tech window (Figure 9) includes necessary skills for regular users to stay secure. Such skills are, for example, choosing a strong password and recognising a malicious e-mail [54]. Those much less technical aspects are connected to terms such as security awareness [37], cybersecurity culture [33], and cyber hygiene [54].

One option for checking security behaviour is using self-assessment and knowledge tests. The disadvantage of this approach is that while knowledge and behavioural self-assessment are generally found to be correlated with the actual security behaviour, they could still measure one thing (security behaviour) by another thing (questionnaires).

Phishing tests can be highly realistic way of simulating an actual cyber-attack, because it is possible to use the content of actual phishing e-mails and only slightly modify the malicious links and attachments so that they would pose no real threat to the targeted people and their organisation. Simulated phishing attacks are also one of the most scalable ways to test the human behaviour.

Sending socially engineered messages to employees and monitoring their reaction seems like a fairly straightforward and simple process. Nevertheless, it should be carried out with caution. Carelessly executed phishing tests might lead to various undesired outcomes, from a situation where users do not click on any legitimate links anymore to legal prosecution. Several papers analyse different aspects of phishing, how to identify it and fight against it (Publication IV). However, there is not much discussion on the particular methods nor the specific process of conducting simulated phishing experiments.

5.2.2 Objectives of a solution

Several design choices can easily be overlooked when designing a phishing simulation test with no specific process. The aim of the case study was to describe the process of carrying out a phishing experiment. Not only technical, but also legal and ethical aspects should be included in the process. Also, the suitability for phishing simulations for both industry and classroom context should be evaluated.

5.2.3 Design and development

The phishing process was developed iteratively in discussions with practitioners who are responsible for information security in their organisation. Also, specialists from university and national computer emergency response team were used to get additional perspectives. The resulting 10-step process (Figure 10) is described in detail in Publication IV. While all of the 10 steps are important, the crucial one is getting permission from an authorised party—otherwise, the experiment can easily be considered illegal in many contexts.

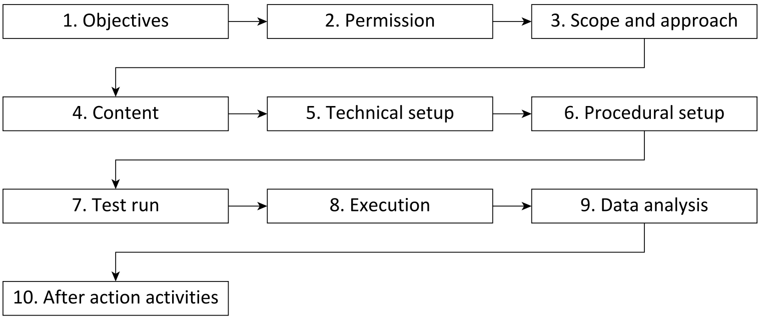


Figure 10 – General process for conducting a successful phishing experiment

5.2.4 Demonstration

The method was implemented in the embedded case study [80] comprising several phishing experiments. Phishing experiments were carried out in 5 different organisations including both private and public sector, some of which are considered part of the national critical infrastructure. Crossover design was used to better understand the impact of different types of phishing email content. After sending simulated phishing emails to employees, individual quasi-structured interviews were held with selected participants. Qualitative side of the experiment helped to better interpret the quantitative results.

5.2.5 Evaluation

Comparative analysis of different organisations showed that the results of the phishing test can be highly context dependent. For example, let us analyse a (hypothetical) case of a fraudulent email inviting to register for the birthday party of the organisation. The low amount of people reacting to this email might be due to several factors. People might be technically savvy and recognise the phishing attempt as fraudulent or just be overloaded with emails and skip the phishing email due to lack of time. People might also be uninterested in the particular type of event mentioned in the email. A similar range of factors influences the reaction to any type of email. Therefore, a phishing test should only be used in a very carefully planned way considering the context of the particular organisation and not drawing any generalisations from the results.

In the classroom education context, the situation is even more complex. In the same organisation, employees are usually required to follow particular security guidelines and rules. However, in the classroom context this assumption does not hold as students of different backgrounds can take part of the same course. Some students may be using their work computer with enforced security policies. Working students might have even participated in a security awareness training. Other students use their own device and might have never heard of any organisational security policy.

In conclusion, it can be said that phishing tests are difficult to conduct in a positively meaningful way. Based on conducted experiments, the phishing simulation itself is not very suitable for classroom education. It can be a good topic for discussion though. While conducting a phishing campaign is a relatively easy and well scalable way to get some insights into the security posture of an organisation, it has serious downsides for using in a classroom context. As an example, a deceiving email that seemingly originates from an authoritative source such as the professor or the university administration can damage the trust towards this source. Nevertheless, the described 10 steps (Figure 10) help to keep in mind the various phishing experiment aspects, that otherwise might be overlooked.

5.3 OSINT exercise (quadrant 4)

5.3.1 Problem identification and motivation

The term Open Source Intelligence (OSINT) originates from military usage and is defined here as the gathering and processing of publicly available information. As increasing volumes of information is publicly available on the Internet, it is important to have the skills to analyse the available data (Publication V). Reconnaissance is the first step in various cybersecurity attack and defence processes. Knowing how to conduct OSINT investigation is an essential part of reconnaissance. Some sources estimate that OSINT can provide between 80% to 98% of all information that a government or private sector organisation needs to know for making informed decisions [34]. However, simulating OSINT tasks in a virtual lab environment presents some inherent technical difficulties.

Many cybersecurity exercises utilise capture-the-flag (CTF) formats where participants are challenged to find or protect specific information (flags). While it is possible to generate individual flags for each participant in a normal cybersecurity exercise context, it is much more difficult to do it for OSINT tasks (Publication V). For example, an OSINT task may include geo-locating an image posted on social media by examining the metadata generated when it was created. However, it is difficult to ensure that after the first student has discovered this information that it is not then shared among their classmates. For that reason, OSINT competitions like the TrendMicro OSINT Challenge[2] are generally not suitable for a classroom education context.

An alternative method is to create separate social media profiles with unique material and custom metadata for each exercise participant. However, this approach may also be not feasible because it does not scale well to large numbers of students. A compromise solution may be to simulate public web sites such as Google and Twitter in the exercise environment. However, this too is problematic as popular web sites tend to be very complex and data intensive. Additionally, the learning experience from engaging with an artificial, simulated public web site such as Google search may be very different from the experience of engaging with the live Internet. (Publication V)

5.3.2 Objectives of a solution

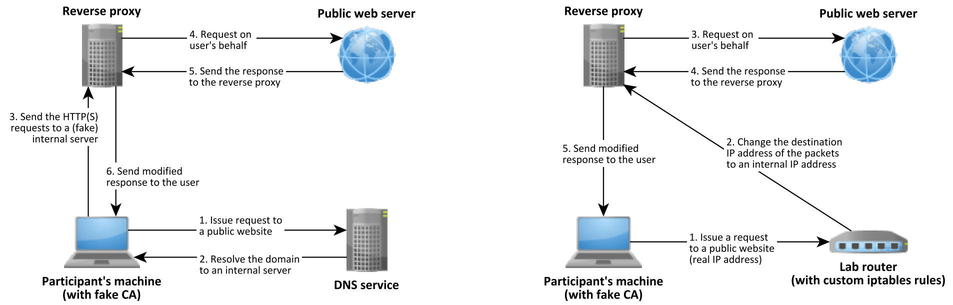
The method for designing virtual cybersecurity exercises related to OSINT competencies should fulfil the following requirements:

- create tasks that are personalised for each participant, so that one participant cannot simply copy the answer from another;
- realistic simulation of publicly accessible web platforms such as Shodan, Facebook, and Google;
- low impact to other people who are visiting the actual web platforms that are simulated.

[2] https://resources.infosecinstitute.com/trend-micro-osint-challenge

5.3.3 Design and development

Publication V presents a technical cybersecurity-related exercise (4th quadrant of the Cybersec-Tech window described in Publication II) targeting OSINT-related skills. While various technical cybersecurity exercises have been created to teach hands-on hacking skills, they usually use an isolated sandbox environment. Isolation helps to ensure that the investigated malware and offensive cybersecurity tools would not cause any harm in systems outside the lab environment. This kind of containment does not allow for realistic open source information gathering that is otherwise an essential step in most attack scenarios. The described method uses man-in-the-middle attack (Figure 11) to insert flags into actual real-life systems such as Twitter or Facebook. The inserted flags are individual based on participant and only visible through the exercise environment.



Reverse proxy
Public web server
4. Request on user's behalf
5. Send the response to the reverse proxy
3. Send the HTTP(S) requests to a (fake) internal server
6. Send modified response to the user
1. Issue request to a public website
2. Resolve the domain to an internal server
Participant's machine (with fake CA)
DNS service
Reverse proxy
Public web server
3. Request on user's behalf
4. Send the response to the reverse proxy
2. Change the destination IP address of the packets to an internal IP address
5. Send modified response to the user
1. Issue a request to a public website (real IP address)
Participant's machine (with fake CA)
Lab router (with custom iptables rules)

Figure 11 – Modifying user traffic using DNS Cache Poisoning (left) and using iptables (right)

Domain Name System (DNS) cache poisoning or DNS spoofing is one of the most prominent DNS attacks. A DNS resolver stores an invalid record for the DNS queries, which causes the domain to be mapped to an alternative IP address. The virtual machines of the exercise labs were in full control of the lab developer. This meant that it was possible to deliberately resolve the DNS queries of users to an internal server instead of their original IP address. An example script illustrating how the implementation of this technique was created in the OSINT lab is shown in Appendix 3. The left part of Figure 11 illustrates a summary of this technique.

When DNS poisoning is used to map the domains to an internal IP address, the user's machine receives an alternative IP address instead of the real address. If users investigate the IP address accessed, DNS poisoning prevents them from accessing the genuine mapped IP address of the domain and prevents its further examination. To address this limitation, iptables are used to change the destination of packets and eventually redirect HTTP and HTTP traffic to a different IP address. Redirecting users without DNS poisoning allows users to see the real IP address of the destination via regular DNS queries, but the lab server redirects the traffic to an internal web server without notifying the users. The right part of Figure 11 shows a diagram illustrating how modifying user traffic using iptables works.

5.3.4 Demonstration

In the initial implementation phase, twenty-six different flags were created for the users to find. These flags were inserted in pictures, app profiles, emails, 10 public web sites, and 4 in-house developed websites to simulate the sites that are owned by the hacker. Since flags are individual strings, it is easy to remove or add new flags to the game in a modular manner. This makes it possible to change the location of the flags or customize them more efficiently. Five master (graduate) level students not participating in the main course were asked to test the initial lab and give their feedback. Based on the feedback, the lab guides were revised to be easier to understand and more specific and a dynamic graph visualising the discovered and undiscovered flags was added. The OSINT lab was then used by 94 students enrolled on a cybersecurity course. Students were divided into two groups – both having one week to complete the lab. However, one group was asked to participate immediately after an introductory lecture and the other one week later. Publication IV gives further details about the design and implementation of the OSINT simulation exercise.

5.3.5 Evaluation

Overall there were 26 flags in the OSINT lab. The number of flags found by the participating students was between 1 and 22. Students in the second group (one week later than the first group) were able to find many more flags than the first group. One possible reason could be that students were being given hints and the location of the flags by the earlier group. Another interesting observation is that in the first day of the exercise, students of the second week were able to find four times the number of flags found in the first day by the first week's group. It took more than three days for the first group to find the same number of flags that the second group found during the first day. This raised the question whether the sudden increase at the beginning of the second group's exercise was because some students shared how to find the flags with other students. Another observation concerns the flag that was found by sending a real email to an address that could only be found during the exercise. It took eight days after the lab was initially given to students for the first email to be sent. However, after the flag was successfully retrieved six different correct emails rapidly followed providing another indicator of how fast the methods for receiving flags were conveyed amongst the students.

To further investigate the information sharing aspect, an additional e-mail was sent out to students after the end of semester asking for their honest view on the homework solving process. 6 students replied. Out of those, 3 students said that they did the exercise independently and did not notice any information sharing. Another 3 students noticed some information sharing – mostly generic ideas shared in smaller groups. One of the students who noticed information sharing mentioned that the information offered was not very helpful. Another student mentioned that the difference in results might have been caused by varied study load in different weeks. Two cases were found where one student tried to submit the flag of another student despite the significant penalties that were advertised to the students beforehand. All connected individuals got penalty points causing a lower grade for the course. The fact that students tried to cheat even fully knowing that they will very likely be discovered and punished shows how important it is to construct assignments that are personalised.

At the end of the course, students were given a questionnaire to assess how useful, interesting, and difficult the lab was. The Questionnaire consisted of three 6-point Likert scale (from 0 to 5) questions. Out of 94 participants, 85 students answered all of the evaluation questions. The lab was rated as "quite difficult" (average rating 4.22 and 4.33 respectively for group 1 and group 2), "rather interesting" (3.33 and 3.65) and "rather useful" (3.11 and 3.26). Students in the second group found the OSINT lab to be of a similar difficulty to the first group, but slightly more interesting and useful than the first group. The first group did not benefit from the experience and shared knowledge that was available for the second group. It is considered possible that further guidance and more thorough preparation for the labs would have increased the positive feedback.

Comparing the results of each week's performance revealed that students actively communicate with each other. This was further confirmed by anecdotal evidence from students' feedback. Therefore, it demonstrates the importance of implementing anti-cheating mechanisms and having individual flags.

To conclude, the developed method satisfied the set requirements. Using man-in-the-middle attack, it was possible to add individual (personalised) content to publicly available web sites. As the added content was visible only through the exercise environment, it had no impact to normal web traffic outside of the lab.

5.4 Measuring cyberethical behaviour (quadrant 1)

5.4.1 Problem identification and motivation

While there is a lot of discussion about the importance of soft skills, they are still rarely measured. This case study focuses on cyberethical behaviour. Despite the vast array of complex ethical dilemmas in the field of modern computer science, there seems to be surprisingly little focus on cyberethics in educational programmes. While there are some initiatives for creating interactive ethics courses [11] and introducing ethical dilemmas into social science courses [84], the general approach tends to stay rather generic and theoretical.

5.4.2 Objectives of a solution

The following case study explores the feasibility of incorporating additional so-called "soft" measurements to technical simulations. The concrete objective is to add cyberethical measurements to an existing virtual simulation and evaluate the feasibility of such process.

5.4.3 Design and development

Ethical measurements were designed to be added to a computer science homework as an additional element (similar to adding a cherry to a cake as shown in Figure 12). Figure 12 illustrates how virtual exercises (or simulations) usually consist of the exercise environment (virtualisation platform), specific content (virtual machine, network topology, tasks), and narrative (storyline). Narrative can be communicated both in the exercise environment or outside of it (e.g., instructor describes the situation in person or by e-mail).

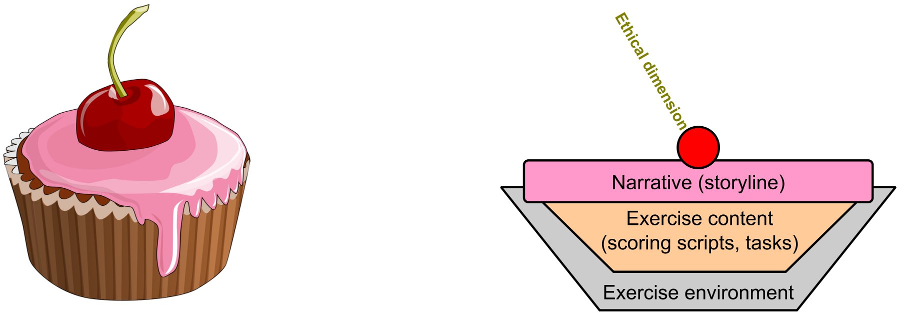


Figure 12 – Illustration of the metaphor of adding ethical dimension to an exercise like adding a cherry to a cupcake. Realistic cupcake with the cherry on the left and exercise diagram on the right. Best viewed in colour.

Publication VI describes how almost any computer science homework can be enriched by adding an ethical task or dilemma. It argues that it is essential to teach students ethical decision making in a safe (simulated) environment.

5.4.4 Demonstration

A case study was conducted where ethical tasks were added to three technical virtual simulation exercises (Publication VI). The case study was carried out as a part of a bachelor (undergraduate) level introductory cybersecurity course for IT-students. 86 students gave their consent to take part in this experiment, and the following analysis is based on their data. There were 72 male (84%) and 14 female participants (16%). The 86 participants were randomly allocated into two groups: group X (45 students) and group Y (41 students). The case study contained three different virtual labs with added ethical tasks, a lecture on the topic of cyberethics, and a questionnaire sent to students in the very beginning and end of the study. Figure 13 illustrates the general course flow regarding the questionnaires, lecture, and virtual labs with the added ethical component.

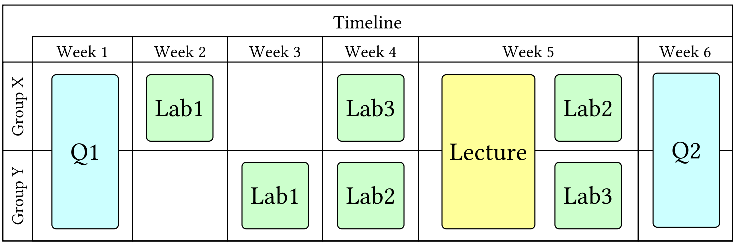


Figure 13 – General timeline of the experiment measuring cyberethical behaviour. Q1 and Q2 stand for the use of questionnaire before and after different labs and lecture. Lab1, Lab2, and Lab3 signify remotely accessible virtual labs with an added ethical component.

A questionnaire was developed to measure attitude and ethical views of students before and after homework assignments containing ethical aspects and the lecture about ethics. The same questionnaire was used twice: before and after the other activities. The questionnaire was added to the experiment to compare self-reported ethical values of the students with their actual ethical behaviour during the completion of their homework. The questionnaire was compiled using selected parts from the Oxford Utilitarianism Scale (OUS) [46], the Cybersecurity Attitude Scale (CAS) [42], and Human Aspects of Information Security Questionnaire (HAIS-Q) [73]. 6-point Likert scale was used to avoid the neutral choice in the middle of the scale [14].

5.4.5 Adding ethical aspects to homework

There are many ethical aspects that can be added to a technical homework assignment. Adding an ethical aspect to the exercise, means introducing an ethical dilemma to the student, where one behaviour is more convenient (or otherwise beneficial) and the other way is more ethically acceptable. The dilemma is usually part of a fictional storyline connected to the task (e.g. finding an emergency fix to a problem until the responsible specialist would return from a holiday).

Schumann et al. [84] list some key elements for an ethical dilemma including the forced choice between an ethical behaviour and a more profitable (or convenient) one.

In computer science, some topics for ethical dilemmas can be the following:

- using pirated/unlicensed software;
- setting up systems with insufficient privacy controls, and security vulnerabilities;
- (not) asking permission for conducting security testing or data processing;
- fixing vulnerabilities (finding the root cause or just dealing with superficial patching);
- dealing with unethical tasks (e.g., adding malicious or illegal functionality to a system).

In the current case study, the dilemma of asking permission to start a technical task was used. Especially in the field of cybersecurity, having a proper written permission to conduct security testing is what often differentiates legitimate specialists from illegal criminals. As part of the homework storyline, students were informed that it would be polite to ask their (fictional) boss for written permission before starting their technical tasks. If the student chose not to ask permission, then they risked losing points. The probability of losing points depended on the lab. The chance of losing points, when not asking permission, was .00001 in Lab 1, and .001 in Lab 2 and Lab 3. The aim was to find out whether students are more likely to ask for permission if the chance of losing points is higher, than if the chance of losing points is lower. Students who asked permission did not receive any additional points. Therefore, the perceived risk of consequences from unethical behaviour was designed to be very low.

The permission request was to be sent to a given e-mail address and the reply was to be expected in no more than 72 hours. What the students did not know, was that actually the reply was given in 3..24 hours and that (randomly chosen) half of the replies required further action—e.g., solving a CAPTCHA. Further action had the sole purpose to make the ethical behaviour of half of the respondents even more time-consuming and therefore inconvenient. The task of asking permission was designed as an additional inconvenience that would potentially significantly limit the time available for doing the homework assignment.

5.4.6 Evaluation

To measure students' activity in the virtual hands-on labs, they were classified, based on their actions, into six groups (G1..G6):

- G1—Students, who asked permission and did not start the lab before receiving one.
- G2—Students, who asked permission but started the lab before they received it.
- G3—Students, who asked permission, but after receiving answer requiring further action, they did not reply—hence they did not receive permission to start the lab but nevertheless did so.
- G4—Students, who asked permission and received permission, but did not start the lab (did not get results).
- G5—Students, who did not ask permission but started the lab.
- G6—Students, who did not ask permission nor started the lab.

Table 1 – Student behaviour per lab

	G1	G2	G3	G4	G5	G6
X-Lab1	9	15	1	2	5	7
Y-Lab1	0	13	5	0	14	2
X-Lab3	1	5	2	20	1	10
Y-Lab2	2	16	4	2	9	1
X-Lab2	6	16	4	3	7	3
Y-Lab3	10	11	4	0	8	1

Table 2 – Cronbach's alphas for pre- and post-questionnaire categories

	IB	IH	PA	PV	HAIS-Q
Q1	0.68	0.75	0.77	0.86	0.48
Q2	0.72	0.74	0.89	0.90	0.75

Spearman's rank-order correlation was calculated between questionnaires and behaviour in the labs. No significant correlation was observed.

Table 2 lists Cronbach's alphas per questionnaire category for both pre-questionnaire (Q1) and post-questionnaire (Q2). Cohen et al. [22] consider alpha values under 0.60 unreliable. The results indicate that in general the questionnaire was a good fit to measure ethical views and cybersecurity attitude. When comparing to other subscales the improvement in HAIS-Q items subscale show unexpectedly big change from 0.48 in pre-questionnaire to 0.75 in the post-questionnaire. This indicates a likely change in the mindset of students. It might be caused by the increase in the understanding of questions or students thinking more carefully, when answering.

Interestingly, no significant Spearman's rank-order correlation was observed between questionnaires and behaviour in the labs. That could indicate that one of the measurements (either the questionnaire or behaviour in the simulated ethical dilemma) is invalid or that the answers to the questionnaire are not strongly connected to the actual behaviour. Criticism towards behavioural models such as Theory of Planned Behaviour [87] seems to support the possibility that the reported behaviour (or attitude and other parameters) based on questionnaire results might not be strongly connected to the actual behaviour. Nevertheless, it could be argued that the student behaviour in the gamified classroom context is also not completely representative of the actual behaviour (e.g., due to Hawthorne effect [62]). While work sample testing has proven to be valid predictor of job performance [79], the interpretation of ethical behaviour measurements in gamified simulations still needs further research.

For group Y, it is visible that the results improved with each lab. Whereas in the first lab only 38% of students got permission, in the final lab, the number had increased to 62%. Also, the number of students, who waited before starting the lab for permission (G1) increased. In the first lab (Y-Lab1) not a single student waited before receiving permission (G2), in the second lab (Y-Lab2) 6% waited for permission. In the third lab (Y-Lab3), after the cyberethics lecture was held, 29% of the students waited before starting the lab. This gives some indication that cyberethics lecture had a positive effect on the cyberethical behaviour of students.

Contrastingly, the results for group X were different. The number of students who got permission in the first lab (X-Lab1) was 67%, in the second lab (X-Lab3) this percentage stayed the same and then decreased to 64% in the third lab (X-Lab2). The reason for this might be that in group X were more students who already knew at the beginning of the experiment about cyberethics behaviour. However, this is not very likely, because the difference between first lab results is big, but in pre-questionnaire, the group difference is not so evident. The gap between the groups X and Y comes clearly out from the results of the first lab. When in group X 67% of the students got permission in the first lab, in group Y 38% of the students got permission. Therefore, the reason for this might be the different timing. Group X did the first lab right after the pre-questionnaire, compared to group Y, who did the first lab a week later, and it might have affected the result of the lab (especially considering potential unsupervised and uncontrollable communication between the groups).

5.5 Summary

This section explored three cybersecurity-related simulations from three different Cybersec-Tech window quadrants.

The phishing simulation was found to be complex and very context dependent. The analysis of phishing simulation suggests that it is not well suited for classroom education in the same form as it is conducted in organisations. However, the topics related to phishing should be discussed. Potentially, a virtual simulation could be designed where participants would need to analyse phishing emails or distinguish them from others.

The OSINT exercise showed how the concept of man-in-the-middle attack can be used to create a scalable and realistic environment for teaching OSINT-related skills. A similar method could also be used outside of cybersecurity context because there are many jobs where the need to find publicly available information is equally important. In cybersecurity education context, the OSINT method opens up a lot of opportunities for designing engaging exercises.

The method for adding cyberethical measurements explored the opportunities to add additional non-technical measures to virtual simulations. The results support using this method and finding ways how to adapt it to different situations. The successful implementation of the method also shows that virtual labs are highly flexible for evaluating a variety of competencies.

Together, the described cases demonstrate the flexibility of cybersecurity simulation exercises. The current section has given specific examples illustrating how the competency-driven exercise design (described in section 4) could look like in practice. The next section shows how mapping various exercises to a competency framework can give additional insights in a classroom context. The results from competency-driven simulation exercises and knowledge tests are combined and mapped to NIST NICE framework in an undergraduate level course.

6 Integrated skill evaluation in classroom context

This section builds upon the Publication VII.

6.1 Problem identification and motivation

To address the shortage of cybersecurity competencies in the job market, the educational processes must become more effective. The final grade for a course is considered as a student's performance indicator and practical exercises often make up a fixed percentage of the final grade. However, grade inflation occurs in many countries [68]. Universities are faced with the dilemma of whether it is more important to provide the maximum number of students with a required set of baseline competencies or to ensure a normal distribution of grades. Normal distribution of grades would help better compare the proficiency levels of different students by avoiding a greater concentration of grades at the upper tail [68]. At the same time, allowing the performance of others to influence individual grades would further complicate the interpretation of the result as it becomes highly context-dependent.

In general, the grading of courses and the scoring of cybersecurity exercises suffer from a similar shortcoming with the lack of a meaningful interpretation of the results (Publication II). While it could be argued that a higher score in a cybersecurity competition and a higher grade in a cybersecurity course both signify a higher proficiency in the topic, a deeper analysis of underlying competencies is often difficult if not impossible. As an example, assume two students both receive a grade 7 (out of 10) in a cybersecurity course. It is usually not possible to deduct from the grade whether one student is more proficient regarding a specific learning outcome than the other. That is because the learning outcomes of the course are rarely evaluated individually. Furthermore, the scores from theoretical knowledge tests and practical skills evaluations are often not differentiated. It is therefore often not possible to see whether a student has demonstrated good results in theory or in practical exercises, or in both.

6.2 Objectives of a solution

The goal is to achieve more systematic overview of individual competencies of each student in a classroom education context. The method for achieving this goal includes the following:

1. Measured competencies should be structured systematically into categories and/or clusters.
2. Meaningful proficiency levels at the higher and lower level tasks should have a clearly defined qualitative difference and not just arbitrary "difficulty level".
3. Differentiation between the evaluation scores of knowledge and skills.

6.3 Design and development

Following the objectives set earlier, the design process comprised the following phases:

1. selecting a suitable cybersecurity competency framework listing relevant knowledge and skills and/or topic areas;
2. selecting a suitable educational taxonomy for meaningful proficiency levels;
3. mapping the tasks in the course to the selected competency framework and assigning the proficiency levels according to the educational taxonomy.

6.3.1 Cybersecurity competency frameworks

While there are multiple initiatives to compile a list of subjects relevant to cybersecurity [72, 35], the most commonly used tend to be the following three. First, a set of curricular recommendations in cybersecurity education (CSEC2017 [16], CSEC2020 [91]), released by a joint task force led by the Association for Computing Machinery and the IEEE Computer Society. Second, the Centers Of Academic Excellence In Cyber Defense Education (CAE-CD) [65] created in the USA by the National Security Agency and the Department of Homeland Security. To be recognised as CAE, a university must have a designated curriculum including a required set of knowledge units [65]. Third, the NIST National Initiative for Cybersecurity Education Cybersecurity Workforce Framework (NICE framework) [66] listing the knowledge, skills, and abilities that are needed for cybersecurity-related jobs.

In this research, the NICE framework [66] was used. This was because it is widely used both in research and in private sector. It also provides an extensive list of knowledge and skills suitably categorised into cybersecurity work roles enabling an assessment of student performance from a perspective of potential future work roles. The NICE framework is comprised of the following components:

- 7 high-level categories: Securely Provision (SP), Operate and Maintain (OM), Oversee and Govern (OV), Protect and Defend (PR), Analyze (AN), Collect and Operate (CO), Investigate (IN).

- 33 specialty areas: Risk Management (RSK), Software Development (DEV), Systems Architecture (ARC), Technology R&D (TRD), Systems Requirements Planning (SRP), Cyber Investigation (INV), Test and Evaluation (TST), Systems Development (SYS), Data Administration (DTA), Knowledge Management (KMG), Customer Service and Technical Support (STS), Network Services (NET), Systems Administration (ADM), Systems Analysis (ANA), Legal Advice and Advocacy (LGA), Training Education and Awareness (TEA), Cybersecurity Management (MGT), Strategic Planning and Policy (SPP), Executive Cyber Leadership (EXL), Program/Project Management (PMA) and Acquisition, Cybersecurity Defense Analysis (CDA), Cybersecurity Defense Infrastructure Support (INF), Incident Response (CIR), Vulnerability Assessment and Management (VAM), Threat Analysis (TWA), Exploitation Analysis (EXP), All-Source Analysis (ASA), Targets (TGT), Language Analysis (LNG), Collection Operations (CLO), Cyber Operational Planning (OPL), Cyber Operations (OPS), Cyber Investigation (INV), Digital Forensics (FOR)

- 52 work roles: Software Developer, Data Analyst, Program Manager, Vulnerability Assessment Analyst etc…

Each high-level category comprises several specialty areas. Each specialty area contains multiple work roles. Each work role is connected to specific knowledge, skill, ability, and task items.

6.3.2 Educational taxonomies

Several educational taxonomies have been developed to provide a shared language for describing learning outcomes and performance in assessments. Of these, the revised Bloom's taxonomy [2] is widely used. Fuller et al. [31] thoroughly analysed various taxonomies in computer science context and propose a two-dimensional adaptation of Bloom's taxonomy—the matrix taxonomy. Niemierko's "ABC" taxonomy [67] as described by Strachanowska [90] and Fuller et al. [31] was chosen for this study. It aligns well with the concept of distinguishing knowledge from skills and at the same time limits the complexity of proficiency levels. A more complex taxonomy would require a deeper analysis of the particular context of specific learning outcomes and the classification can become problematic [31]. Niemierko's "ABC" taxonomy is generally aligned to Bloom's taxonomy by just combining the three highest Bloom categories into one. It is also somewhat similar to the single-loop and double-loop learning [3]. The single-loop signifies the repetition of the same strategy to solve a problem and double-loop emphasises critical self-reflection to generate creative solutions to problems [47].

6.3.3 Mapping the tasks to competencies and proficiency levels

To map the the tasks to competencies and proficiency levels, 15 knowledge tests and 12 homework assignments of the course were analysed. Knowledge tests were divided into questions and each question was mapped to the relevant knowledge unit of the NICE framework. In this study, all 589 knowledge and 365 skill items listed in the NICE framework master KSA (knowledge, skills, abilities) list were used.

To map the 111 questions used during the course to 589 knowledge items in NICE framework, an initial filtering of relevant knowledge items was conducted. Initial filtering identified 75 knowledge items relevant to the course. Each question was assigned difficulty level of 1 or 2 based on Niemierko's "ABC" taxonomy categories A and B (see Table 3). An example of a level 1 question would be *"What does the letter A in CIA triad stand for?" (Right answer: availability.)* An example of a level 2 question would be *"Mary adds her digital signature to her essay. Which security aspect does it help to protect?" (Right answer: integrity.)* Answering correctly the level 2 question results in a proficiency point in both level 1 and level 2.

The tasks from the practical homework assignments were connected to relevant skills in the NICE framework. For establishing proficiency levels, each homework was assigned with thresholds for achieving difficulty levels 1 and 2 based on Niemierko's "ABC" taxonomy categories C and D. For example, the tasks in the reverse engineering exercise were connected to NICE skills *S0087—Skill in deep analysis of captured malicious code (e.g., malware forensics)* and *S0131—Skill in analysing malware*. In this lab, students had to analyse executable files with increasing complexity. There was 1 executable program in Java, and 6 executables written in C. Solving the reverse engineering task for the executable written in Java gave students 1 point. Solving the reverse engineering tasks connected to executables in C, gave 0.5 points per task. The total possible number of points from the homework was 4 and the maximum proficiency level to related competencies was level 2. The reverse engineering virtual lab had the threshold of at least 1 point (out of 4) for achieving level 1 proficiency and the threshold of at least 3 points (out of 4) for achieving both level 1 and level 2 proficiency.

Difficulty levels determining the proficiency are summarised in Table 3. The summary of the designed process is given in Figure 14.

6.4 Demonstration

The case study was conducted in an introductory undergraduate level cybersecurity course with 106 participants. Performance scores were gathered from 15 knowledge tests that took place in Kahoot[3] environment during weekly lectures and 12 homework assignments. Every test and homework assignment contributed points towards the final grade. There were three types of homework assignments: 1 essay (evaluated by peers based on de-

[3] https://kahoot.com

Table 3 – Difficulty levels based on Niemierko's "ABC" taxonomy [67] categories

Difficulty	Competency component	ABC taxonomy category
1	knowledge	A. Remembering the knowledge
2	knowledge	B. Understanding the knowledge
1	skills	C. Application of knowledge in typical problem situations
2	skills	D. Application of knowledge in unfamiliar problem situations

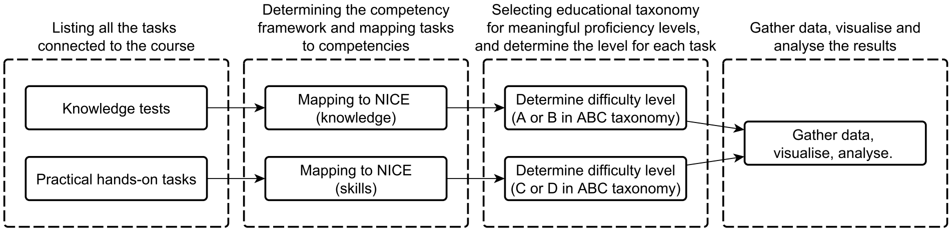

Listing all the tasks connected to the course
Determining the competency framework and mapping tasks to competencies
Selecting educational taxonomy for meaningful proficiency levels, and determine the level for each task
Gather data, visualise and analyse the results
Knowledge tests
Practical hands-on tasks
Mapping to NICE (knowledge)
Mapping to NICE (skills)
Determine difficulty level (A or B in ABC taxonomy)
Determine difficulty level (C or D in ABC taxonomy)
Gather data, visualise, analyse.

Figure 14 – Competency mapping process. Sections with dashed lines represent the generic steps for the described competency mapping method and sections with solid lines represent the steps taken during the current case study.

tailed criteria), 3 traditional assignments (manually graded report and/or software solution), and 8 virtual labs in a remotely accessible cloud-based environment (automatically scored).

Knowledge was assessed based on a binary evaluation of Kahoot test answers (right or wrong) of each student. Each right answer gave 1 or 2 points according to its difficulty level (Niemierko's "ABC" taxonomy [67] categories A and B), contributing to the total score of related NICE knowledge units. Skills were assessed based on homework assignments. Results from home assignments contributed to grades (max 4 to 10 points depending on homework) and also to skill evaluations that were based on competency mapping to NICE framework.

6.5 Evaluation

In general, the method used fulfilled the set criteria. Measured competencies were mapped to NICE framework providing a well-structured systematic overview. Proficiency levels using the guidelines of Niemierko's "ABC" taxonomy [67] provided a good balance between simplicity and meaningful interpretation. Niemierko's "ABC" taxonomy also enabled knowledge and skill evaluation scores to be differentiated.

The mapping process itself provided useful insights regarding the course. From 589 knowledge items of the NICE framework, only 75 were considered to be relevant to the existing course. The rest of the knowledge items provide a good basis for considering additional topics that could be included in the course in the future. Also, out of the 75 relevant knowledge items, only 44 were actually measured by the quiz demonstrating a significant gap between the topics covered by the lectures and later knowledge measurements. For example *K0110—Knowledge of adversarial tactics, techniques, and procedures* was mentioned in several lectures, but never included in any test.

Additionally, the mapping provided the opportunity to see which NICE specialty areas were used the most during the course assessments. For example, it can be seen in Figure 15 that the Securely Provision (SP) category had a strong focus during the course while Investigate (IN) category was covered less.

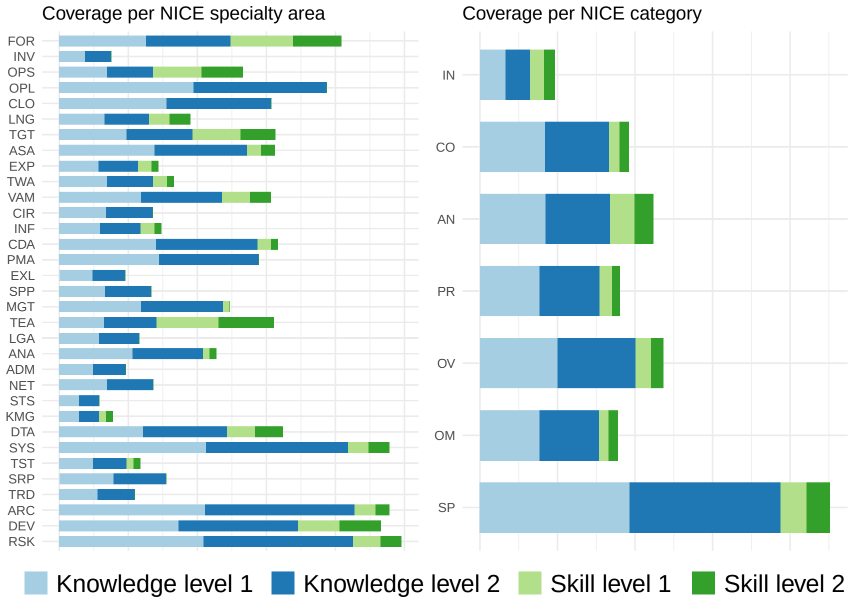


Figure 15 – Maximum possible results from course assessments categorised into NICE framework knowledge and skills and afterwards grouped into NICE specialty areas (on the left) and NICE categories (on the right).

Nevertheless, it should be noted that the NICE framework is a high-level framework and does not go into details regarding skills and knowledge items. Therefore, the mapping might not always adequately reflect the coverage of the concept. For example, consider a student who has completed a technical task mapped to NICE skill *S0130—Skill in writing scripts using R, Python, PIG, HIVE, SQL, etc*. From seeing the skill, it is not possible to deduct what programming language was used. Furthermore, the level of proficiency is difficult to determine as the scope of “writing scripts” can be understood differently. However, some sort of ambiguity is inevitable because defining a competency with infinite accuracy is not feasible in most practical contexts.

Mapping the results of students provides further insights on their competencies. As an example, Figure 16 compares the results of Alice and Bob (pseudonyms of real students) based on NICE categories and specialty areas. Figure 17 shows the same results using the normalised scores for each of the 4 competency proficiency levels (knowledge level 1 and 2; skill level 1 and 2). Alice and Bob represent the results of two students whose overall scores were almost identical, and their final course grade was the same. However, a detailed look into the performance of each specific area can help find potentially interesting and suitable areas for analysis. For example, it can be seen that Bob achieved good results (skill level 2) in Cybersecurity Defense Analysis (CDA), and Alice in Exploitation Analysis (EXP). While these insights are based on initial data and should be carefully analysed further, they illustrate how competency mapping can provide relevant ideas for future skills development.

Overall, the mapping method fulfilled its objectives and provided multiple additional insights as well as ideas for future research.

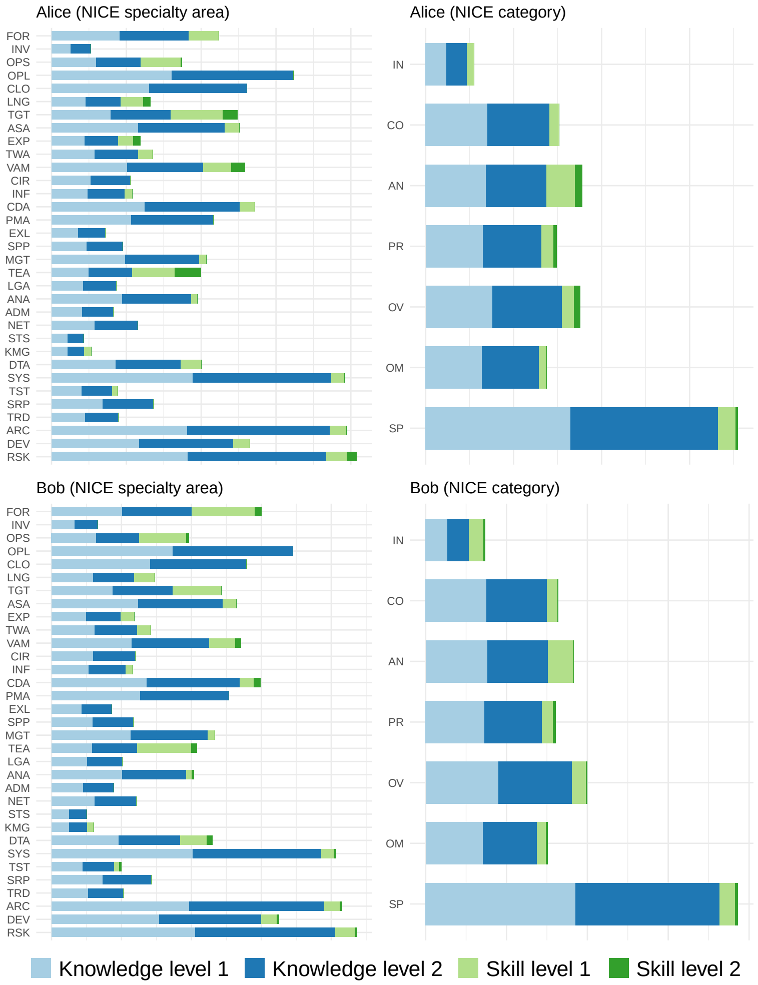


Figure 16 – Comparing the summarised results from two students using NICE framework specialty areas (on the left) and NICE framework high-level categories (on the right). Please refer to NICE framework documentation [66] for detailed list of specialty areas, and categories.

6.6 Summary

The earlier sections laid the foundation for creating competency-driven cybersecurity simulation exercises and provided examples illustrating their use and flexibility. This section showed how competency-driven simulation exercises can be used in a classroom context. Mapping results from various tasks to the NICE framework gave a good overview of individual competences and it also helped to understand and plan the measurement of learning outcomes of the course in general.

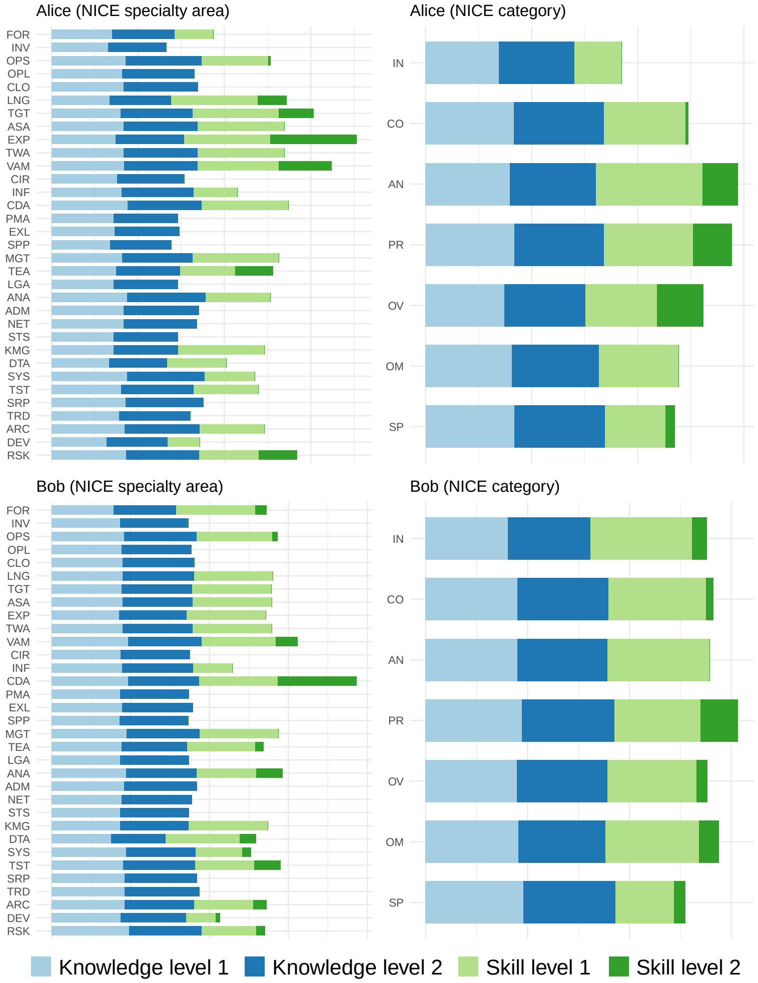


Figure 17 – Comparing the normalised and summarised results from two students using NICE framework specialty areas (on the left) and NICE framework high-level categories (on the right). Please refer to NICE framework documentation [66] for detailed list of specialty areas, and categories.

7 Conclusions and future work

This research contributes to cybersecurity education by providing new insights into competency-driven simulation exercises. More specifically, various novel virtual simulations were created to demonstrate their capability to evaluate a diverse set of cybersecurity-related competencies. Using those virtual simulations as case studies, processes for creating competency-driven cybersecurity exercises were formalised. This can provide guidance for creating effective cybersecurity exercises and therefore help provide more people with relevant cybersecurity competencies.

To describe how to effectively evaluate cybersecurity-related competences through simulation exercises, four research questions were presented in section 1.3 and answered throughout this thesis. Nevertheless, for a simplified overview, the generic answers are also given in the following list.

- **RQ 1: How to categorise cybersecurity-related competences?**
  Cybersecurity encompasses technical concepts ranging from computer science to various human-related psychological topics of social sciences. In general, it is possible to categorise cybersecurity-related competences according to the Cybersec-Tech window (Publication II) based on how technical and how related to cybersecurity they are. This kind of generic classification stresses the importance of both technical and non-technical competencies and makes it easier to have a balanced approach to cybersecurity education.

  More specifically, it is possible to categorise the competencies according to a job description. Identifying required competencies for specific job roles helps to keep the cybersecurity education or training aligned with the expectations in the job market. Furthermore, mapping competency evaluations to specific job profiles enables the potential individual to match to the most appropriate job role.

- **RQ 2: What kind of metrics relevant to skills evaluation can be gathered using virtual labs?**
  The most relevant metric tends to be task completion. It should be noted though, that measuring task completion is not always a straightforward process. If an offensive task has the goal of finding some specific confidential information from the targeted system, it is relatively easy to control whether this objective was achieved. Evaluating defensive competencies tends to be more complex. That is because in cyberdefense, the goal is often to protect the confidentiality, integrity, and availability of the system at the same time. If an attack targeting some confidential data fails, but results in the system becoming not functional or not available at all, then the defense is rarely considered successful. Therefore, it is often essential to consider and measure multiple aspects of the situation before concluding that the task was completed successfully.

  Also, time-related metrics can be useful. For example, even if the cyberattack was successful and the system taken down, it is usually possible to restore at least some of its functionality and/or lost data. The fast recovery of business-critical systems is crucial for real-life organisations. Therefore, it is logical to also assess and evaluate the reaction speed in the cybersecurity simulations.

  In more advanced exercises, participants have usually different ways how to achieve their objectives. For example, an attacker might find a vulnerability in the web application or instead exploit the underlying operating system of the target server.

A defender might patch a web application vulnerability in the source code or configure the firewall to block related attack opportunities. When there are multiple ways to achieve a similar result, it is important to establish measurements capable of distinguishing not only the end result but also the way it was achieved.

- **RQ 3: What kind of cybersecurity-related skills can be evaluated with virtual labs?** Various cybersecurity-related skills can be evaluated with virtual labs. Virtual labs are well suited for evaluating individual technical skills. It is usually possible to simulate technical cybersecurity tasks very realistically. The processes of setting up a web server or configuring a firewall in an exercise environment can be identical those in real life.

  In addition to technical skills, it is also possible to evaluate non-technical skills using virtual simulations. However, the interpretation of the results is not so straightforward as it is with technical tasks. Human behaviour in the simulation context might significantly differ from the behaviour in real life just because people are aware that the exercise is not real. Especially this is the case in classroom education where students might not mind acting in an inadequate or unethical way inside the simulation with the excuse that it is not real. Nevertheless, it is possible to use virtual labs for evaluating various non-technical competencies. As an example, the current thesis described a method for adding cyberethical measurements to a technical virtual simulation.

- **RQ 4: How to design virtual labs suitable for measuring cybersecurity-related skills?** Competency-driven design helps to measure cybersecurity-related skills with virtual simulations. To measure cybersecurity-related skills in a scalable and meaningful way, all the measurements need to be mapped to specific competencies. It is recommended that those competencies follow a structure that enables the interpretation of the final results in a meaningful way. For example, connecting the results from a virtual simulation to competencies of a specific job role enables the suitability of the exercise participant to this specific job to be evaluated. Also, to ensure the comparability of the results throughout different simulations, it would be beneficial to use a widely accepted educational taxonomy. This enables to establish and measure competency levels that are qualitatively different.

This thesis lays the groundwork for promising research.

First, the holistic competency evaluation scheme presented in section 6 could benefit from a supporting information system that helps in gathering and presenting the learners' results. Visualising this kind of learning analytics data in a user-friendly manner is a separate research area in itself. A more detailed view to personal competencies would enable both teachers and students to adapt faster and find suitable educational approaches faster for various frequently changing technological topics. Having fast individual feedback about competencies and related progress can also help in finding more suitable learning paths based on personal goals.

Second, similar competency evaluations could be gathered throughout several courses. Monitoring student performance across different courses would help to adjust the whole study programme according to the measured results. At the moment, curricular design often focuses on covering relevant topics, but measured competency-based performance data can provide new insights about the actual learning processes of students.

Third, the evaluation methods should be further validated. There is abundant research in psychometrics and psychology regarding the validation of assessment instruments. Also, there is recent research looking at developing suitable knowledge assessment instruments for computer science [15]. Still, validating a measuring instrument is a long process. Furthermore, most of the work on educational assessment instruments is focusing on knowledge, not skills or other competencies. Validating skill assessment instruments for cybersecurity, in rapidly changing technological landscape, will surely continue to provide challenging material for research in the future.

Fourth, there could be more collaboration with the industry regarding the competency evaluation. This thesis demonstrated that it is possible to take the NIST NICE competency framework that is used by the industry and integrate it to classroom education. As one of the purposes of education continually tends to be the preparation of the future workforce [9], then it is only natural that the evaluation processes in education would be at least partly aligned to the expectations in the industry.

Last, but not least, this thesis shows the flexibility of virtual simulations for evaluating not only technical skills. There is general agreement on the importance of competencies such as leadership and teamwork, but more work needs to be done to find good ways how to measure them in classroom context. As virtual simulations can quite accurately mimic many realistic work situations, they should include more evaluations of non-technical skills.

In conclusion, virtual simulation exercises provide a fertile ground for future research. This thesis has explored only some ways how simulation exercises could be used to make cybersecurity education more effective and meaningful. Hopefully, the described processes and novel technological solutions inspire more discussions and practical work in making education more relevant, effective, and fun.

Glossary

CAE	National Centers of Academic Excellence established by the National Security Agency of United States of America
CSEC	Curricular guidelines for cybersecurity curricula developed by joint task force led by ACM (publications in 2017 and in 2020). CSEC2017 publication was designed for 4-year institutions. CSEC2020 publication provided the guidelines appropriate at the 2-year level, suitable for institutions such as community colleges in the US.
CTF	Capture the Flag—a game where several teams each have a flag (or other marker) and the objective is to capture the flag(s) of other team(s). Sometimes, in cybersecurity exercises, the goal might also be to plant a flag (e.g., defacement attack) as a proof of successful attack.
DSR	Design science research
DSRM	Design science research methodology. Referring to the formalised process by Peffers et al. [75]
IT	Information Technology
KSA	Knowledge, Skills, and Abilities—that is how KSA is defined by NIST NICE Framework and how it is mostly used in the current thesis. Related work section mentions also some possible alternative definitions for the letter A in KSA.
NICE	National Initiative for Cybersecurity Education (by NIST) and according framework listing work roles and connected KSA
NIST	The National Institute of Standards and Technology that is part of the Department of Commerce in United States of America
NSA	National Security Agency of United States of America
OSINT	Open Source INTelligence - gathering information using publicly available data
SOLO	Structure of the Observed Learning Outcome—educational taxonomy focusing on the content of the learner’s response and its structural relationships.
US or USA	United States of America
XSS	Cross-site scripting—type of computer security vulnerability that enables the injection of malicious client-side scripts to a user’s web browser for execution.

List of Figures

List of Tables

References

[1] M. Adams and M. Makramalla. Cybersecurity skills training: an attacker-centric gamified approach. *Technology Innovation Management Review*, 5(1), 2015.

[2] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock. A taxonomy for learning, teaching and assessing: A revision of bloom's taxonomy of educational objectives: Complete edition, 2001.

[3] C. Argyris. Teaching smart people how to learn. *Harvard business review*, 69(3), 1991.

[4] M. E. Armstrong, K. S. Jones, A. S. Namin, and D. C. Newton. The knowledge, skills, and abilities used by penetration testers: Results of interviews with cybersecurity professionals in vulnerability assessment and management. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 709–713. SAGE Publications Sage CA: Los Angeles, CA, 2018.

[5] M. S. Atkin, D. L. Westbrook, and P. R. Cohen. Capture the flag: Military simulation meets computer games. In *Proceedings of AAAI Spring Symposium Series on AI and Computer Games*, pages 1–5, 1999.

[6] M. Baehr. Distinctions between assessment and evaluation. *Program Assessment Handbook*, 7(1):231–234, 2005.

[7] R. Baskerville, J. Pries-Heje, and J. Venable. Soft design science methodology. In *Proceedings of the 4th international conference on design science research in information systems and technology*, pages 1–11, 2009.

[8] D. C. Berliner. Comment: Educational research: The hardest science of all. *Educational researcher*, 31(8):18–20, 2002.

[9] G. Biesta. Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, 21(1):33–46, 2009.

[10] J. B. Biggs and K. F. Collis. *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. Academic Press, 1982.

[11] J. Blanken-Webb, I. Palmer, S.-E. Deshaies, N. C. Burbules, R. H. Campbell, and M. Bashir. A case study-based cybersecurity ethics curriculum. In *2018 USENIX Workshop on Advances in Security Education (ASE 18)*, 2018.

[12] B. S. Bloom et al. Taxonomy of educational objectives. vol. 1: Cognitive domain. *New York: McKay*, pages 20–24, 1956.

[13] C. Brabrand and B. Dahl. Using the solo taxonomy to analyze competence progression of university science curricula. *Higher Education*, 58(4):531–549, 2009.

[14] J. D. Brown. What issues affect likert-scale questionnaire formats. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 4(1), 2000.

[15] P. S. Buffum, E. V. Lobene, M. H. Frankosky, K. E. Boyer, E. N. Wiebe, and J. C. Lester. A practical guide to developing and validating computer science knowledge assessments with application to middle school. In *Proceedings of the 46th ACM technical symposium on computer science education*, pages 622–627, 2015.

[16] D. L. Burley, M. Bishop, S. Buck, J. J. Ekstrom, L. Futcher, D. Gibson, E. K. Hawthorne, S. Kaza, Y. Levy, H. Mattord, and A. Parrish. *Cybersecurity Curricula 2017*, volume 1. 2017.

[17] K. Cabaj, D. Domingos, Z. Kotulski, and A. Respício. Cybersecurity education: Evolution of the discipline and analysis of master programs. *Computers & Security*, 75:24–35, 2018.

[18] A. Calderón and M. Ruiz. A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education*, 87:396–422, 2015.

[19] M. Carlton and Y. Levy. Expert assessment of the top platform independent cybersecurity skills for non-it professionals. In *SoutheastCon 2015*, pages 1–6. IEEE, 2015.

[20] T. Chothia and C. Novakovic. An offline capture the flag-style virtual machine and an assessment of its value for cybersecurity education. In *2015 {USENIX} Summit on Gaming, Games, and Gamification in Security Education (3GSE 15)*, 2015.

[21] Y.-k. Chou. *Actionable gamification : beyond points, badges, and leaderboards*. Octalysis Group, United States, 2015.

[22] L. Cohen, L. Manion, and K. Morrison. *Research methods in education*. routledge, 2002.

[23] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle. A systematic literature review of empirical evidence on computer games and serious games. *Computers & education*, 59(2):661–686, 2012.

[24] C. Cowan, S. Arnold, S. Beattie, C. Wright, and J. Viega. Defcon capture the flag: Defending vulnerable code from intense attack. In *Proceedings DARPA Information Survivability Conference and Exposition*, volume 1, pages 120–129. IEEE, 2003.

[25] C. Eagle and J. L. Clark. Capture-the-flag: Learning computer security under fire. Technical report, NAVAL POSTGRADUATE SCHOOL MONTEREY CA, 2004.

[26] T. J. Ellis and Y. Levy. A guide for novice researchers: Design and development research methods. In *Proceedings of Informing Science & IT Education Conference (InSITE)*, volume 10, pages 107–118, 2010.

[27] M. Ernits and K. Kikkas. A live virtual simulator for teaching cybersecurity to information technology students. In *International Conference on Learning and Collaboration Technologies*, pages 474–486. Springer, 2016.

[28] M. Ernits, K. Maennel, S. Mäses, T. Lepik, and O. Maennel. From simple scoring towards a meaningful interpretation of learning in cybersecurity exercises. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security: ICCWS 2020*. Academic Conferences and publishing limited, 2020.

[29] P. A. Ertmer and T. J. Newby. Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance improvement quarterly*, 26(2):43–71, 2013.

[30] S. Frezza, M. Daniels, A. Pears, Å. Cajander, V. Kann, A. Kapoor, R. McDermott, A.-K. Peters, M. Sabin, and C. Wallace. Modelling competencies for computing education beyond 2020: a research based approach to defining competencies in the computing disciplines. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pages 148–174, 2018.

[31] U. Fuller, C. G. Johnson, T. Ahoniemi, D. Cukierman, I. Hernán-Losada, J. Jackova, E. Lahtinen, T. L. Lewis, D. M. Thompson, C. Riedesel, et al. Developing a computer science-specific learning taxonomy. *ACM SIGCSE Bulletin*, 39(4):152–170, 2007.

[32] S. Furnell and M. Bishop. Addressing cyber security skills: the spectrum, not the silo. *Computer Fraud & Security*, 2020(2):6–11, 2020.

[33] N. Gcaza and R. von Solms. Cybersecurity culture: An ill-defined problem. In *IFIP World Conference on Information Security Education*, pages 98–109. Springer, 2017.

[34] S. D. Gibson. Exploring the role and value of open source intelligence. In *Open Source Intelligence in the Twenty-First Century*, pages 9–23. Springer, 2014.

[35] J. Hallett, R. Larson, and A. Rashid. Mirror, mirror, on the wall: What are we teaching them all? characterising the focus of cybersecurity curricular frameworks. In *2018 USENIX Workshop on Advances in Security Education (ASE 18)*, Baltimore, MD, 2018. USENIX Association.

[36] J. Hamari, J. Koivisto, H. Sarsa, et al. Does gamification work?-a literature review of empirical studies on gamification. In *HICSS*, volume 14, pages 3025–3034, 2014.

[37] N. Hänsch and Z. Benenson. Specifying it security awareness. In *2014 25th International Workshop on Database and Expert Systems Applications*, pages 326–330. IEEE, 2014.

[38] M. Henri, M. D. Johnson, and B. Nepal. A review of competency-based learning: Tools, assessments, and recommendations. *Journal of engineering education*, 106(4):607–638, 2017.

[39] D. S. Henshel, G. M. Deckard, B. Lufkin, N. Buchler, B. Hoffman, P. Rajivan, and S. Collman. Predicting proficiency in cyber defense team exercises. In *MILCOM 2016-2016 IEEE Military Communications Conference*, pages 776–781. IEEE, 2016.

[40] A. Hevner and S. Chatterjee. Design science research in information systems. In *Design research in information systems*, pages 9–22. Springer, 2010.

[41] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS quarterly*, pages 75–105, 2004.

[42] D. J. Howard. Development of the cybersecurity attitudes scale and modeling cybersecurity behavior and its antecedents. 2018. Master thesis in University of South Florida.

[43] W. Huitt, J. Hummel, and D. Kaeck. Assessment, measurement, evaluation, and research: educational psychology interactive. *Educational Psychology Interactive*, 2001.

[44] K. P. Jantke. Toward a taxonomy of game based learning. In *2010 IEEE International Conference on Progress in Informatics and Computing*, volume 2, pages 858–862. IEEE, 2010.

[45] K. S. Jones, A. S. Namin, and M. E. Armstrong. The core cyber-defense knowledge, skills, and abilities that cybersecurity students should learn in school: Results from interviews with cybersecurity professionals. *ACM Transactions on Computing Education (TOCE)*, 18(3):11, 2018.

[46] G. Kahane, J. A. Everett, B. D. Earp, L. Caviola, N. S. Faber, M. J. Crockett, and J. Savulescu. Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. 2017.

[47] K. Kiili. Foundation for problem-based gaming. *British journal of educational technology*, 38(3):394–404, 2007.

[48] W. Knowles, J. M. Such, A. Gouglidis, G. Misra, and A. Rashid. All that glitters is not gold: on the effectiveness of cyber security qualifications. *IEEE Computer*, 50(12):60–71, 2017.

[49] I. Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, 47(4):2025–2047, 2013.

[50] Y. Li, D. Nguyen, and M. Xie. Ezsetup: A novel tool for cybersecurity practices utilizing cloud resources. In *Proceedings of the 18th Annual Conference on Information Technology Education*, pages 53–58, 2017.

[51] C. A. Lindley. Game taxonomies: A high level framework for game analysis and design. *Gamasutra*, 2003.

[52] S. Lindskog, U. Lindqvist, and E. Jonsson. *IT security research and education in synergy*. Univ., 1999.

[53] C. S. Loh, Y. Sheng, and D. Ifenthaler. Serious games analytics. *Edited by Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler. Cham: Springer International Publishing. doi*, 10:978–3, 2015.

[54] K. Maennel, S. Mäses, and O. Maennel. Cyber hygiene: The big picture. In *Nordic Conference on Secure IT Systems*, pages 218–226. Springer, 2018.

[55] K. Maennel, R. Ottis, and O. Maennel. Improving and measuring learning effectiveness at cyber defense exercises. In *Nordic Conference on Secure IT Systems*, pages 123–138. Springer, 2017.

[56] S. Mäses. Evaluating cybersecurity-related competences through serious games. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, pages 1–2, 2019.

[57] S. Mäses, H. Aitsam, and L. Randmann. A method for adding cyberethical behaviour measurements to computer science homework assignments. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, page 18. ACM, 2019.

[58] S. Mäses, B. Hallaq, and O. Maennel. Obtaining better metrics for complex serious games within virtualised simulation environments. In *ECGBL 2017 11th European Conference on Game-Based Learning*, pages 428–434. Academic Conferences and publishing limited, 2017.

[59] S. Mäses, K. Kikerpill, K. Jüristo, and O. Maennel. Mixed methods research approach and experimental procedure for measuring human factors in cybersecurity using phishing simulations. In *ECRM 2019 18th European Conference on Research Methodology for Business and Management Studies*, pages 428–434. Academic Conferences and publishing limited, 2017.

[60] S. Mäses, O. Maennel, and S. Sütterlin. Using competency mapping for skill assessment in an introductory cybersecurity course. In *Educating Engineers for Future Industrial Revolutions - Proceedings of the 23rd International Conference on Interactive Collaborative Learning (ICL2020)*. Springer, 2020.

[61] S. Mäses, L. Randmann, O. Maennel, and B. Lorenz. Stenmap: Framework for evaluating cybersecurity-related skills based on computer simulations. In *Learning and Collaboration Technologies. Learning and Teaching*, pages 1–13. Springer, 2018.

[62] J. McCambridge, J. Witton, and D. R. Elbourne. Systematic review of the hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology*, 67(3):267–277, 2014.

[63] J. Mirkovic and P. A. Peterson. Class capture-the-flag exercises. In *2014 {USENIX} Summit on Gaming, Games, and Gamification in Security Education (3GSE 14)*, 2014.

[64] R. Nacheva. Prototyping approach in user interface development. In *Second conference on innovative teaching methods (ITM 2017). 28-29 June 2017, Varna*, volume 28, page 78, 2017.

[65] National Security Agency. National centers of academic excellence. `https://www.nsa.gov/resources/students-educators/centers-academic-excellence/`. Accessed February 29, 2020.

[66] W. Newhouse, S. Keith, B. Scribner, and G. Witte. National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework. *NIST Special Publication 800-181*, 2017.

[67] B. Niemierko and W. S. i Pedagogiczne. *ABC testów osiagnieć szkolnych (in Polish)*. Wydawnictwa Szkolne i Pedagogiczne, 1975.

[68] M. Nordin, G. Heckley, and U. Gerdtham. The impact of grade inflation on higher education enrolment and earnings. *Economics of Education Review*, 73:101936, 2019.

[69] D. O'Leary. Expert system prototyping as a research tool. *Applied Expert Systems, North-Holland, Amsterdam*, pages 17–32, 1988.

[70] H. Österle, J. Becker, U. Frank, T. Hess, D. Karagiannis, H. Krcmar, P. Loos, P. Mertens, A. Oberweis, and E. J. Sinz. Memorandum on design-oriented information systems research. *European Journal of Information Systems*, 20(1):7–10, 2011.

[71] E. Ouzounis, P. Trimintzio, and P. Saragiotis. Good practice guide on national exercises, 2009.

[72] A. Parrish, J. Impagliazzo, R. K. Raj, H. Santos, M. R. Asghar, A. Jøsang, T. Pereira, and E. Stavrou. Global perspectives on cybersecurity education for 2030: a case for a meta-discipline. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pages 36–54. ACM, 2018.

[73] K. Parsons, D. Calic, M. Pattinson, M. Butavicius, A. McCormac, and T. Zwaans. The human aspects of information security questionnaire (hais-q): two further validation studies. *Computers & Security*, 66:40–51, 2017.

[74] V.-V. Patriciu and A. C. Furtuna. Guide for designing cyber security exercises. In *Proceedings of the 8th WSEAS International Conference on E-Activities and information security and privacy*, pages 172–177. World Scientific and Engineering Academy and Society (WSEAS), 2009.

[75] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[76] B. E. Penprase. The fourth industrial revolution and higher education. In *Higher education in the era of the fourth industrial revolution*, pages 207–229. Springer, 2018.

[77] P. Pusey, M. Gondree, and Z. Peterson. The outcomes of cybersecurity competitions and implications for underrepresented populations. *IEEE Security & Privacy*, 14(6):90–95, 2016.

[78] P. L. Roth, P. Bobko, and L. A. McFarland. A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, 58(4):1009–1037, 2005.

[79] P. L. Roth, P. Bobko, and L. A. McFarland. A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, 58(4):1009–1037, 2005.

[80] P. Runeson and M. Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131, 2009.

[81] N. B. Ruparelia. Software development lifecycle models. *ACM SIGSOFT Software Engineering Notes*, 35(3):8–13, 2010.

[82] M. Sabin, H. Alrumaih, and J. Impagliazzo. A competency-based approach toward curricular guidelines for information technology education. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pages 1214–1221. IEEE, 2018.

[83] D. Schatz, R. Bashroush, and J. Wall. Towards a more representative definition of cyber security. *Journal of Digital Forensics, Security and Law*, 12(2):53–74, 2017.

[84] P. L. Schumann, T. W. Scott, and P. H. Anderson. Designing and introducing ethical dilemmas into computer-based business simulations. *Journal of Management Education*, 30(1):195–219, 2006.

[85] D. Shoemaker, D. Davidson, and A. Conklin. Toward a discipline of cyber security: Some parallels with the development of software engineering education. *EDPACS*, 56(5-6):12–20, 2017.

[86] M. Smith. *Testing people at work : competencies in psychometric testing*. BPS Blackwell, Malden, MA, 2005.

[87] F. F. Sniehotta, J. Presseau, and V. Araújo-Soares. Time to retire the theory of planned behaviour. *Health Psychology Review*, 8(1):1–7, 2014.

[88] T. Sommestad and J. Hallberg. Cyber security exercises and competitions as a platform for cyber security experiments. In *Nordic Conference on Secure IT Systems*, pages 47–60. Springer, 2012.

[89] C. W. Starr, B. Manaris, and R. H. Stalvey. Bloom's taxonomy revisited: specifying assessable learning objectives in computer science. In *ACM SIGCSE Bulletin*, volume 40, pages 261–265. ACM, 2008.

[90] I. Strachanowska. Taxonomy abc setting operational teaching objectives in the pedagogical process of training english philology students. 1997.

[91] C. Tang, C. Tucker, C. Servin, M. Geissler, M. Stange, N. Jones, J. Kolasa, A. Phillips, L. Piskopos, and P. Schmelz. Cybersecurity curricular guidance for associate-degree programs. Technical report, Association for Computing Machinery (ACM) Committee for Computing Education in Community Colleges (CCECC), 2020.

[92] D. H. Tobey, R. A. Gandhi, A. B. Watkins, and C. W. O'Brien. Competency is not a three letter word a glossary supporting competency-based instructional design in cybersecurity. *Cybersecurity Skills Journal: Practice and Research*, 1, 2018.

[93] R. Vogel et al. Closing the cybersecurity skills gap. *Salus Journal*, 4(2):32, 2016.

[94] R. Weiss, M. E. Locasto, and J. Mache. A reflective approach to assessing student performance in cybersecurity exercises. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pages 597–602. ACM, 2016.

[95] R. Weiss, F. Turbak, J. Mache, and M. E. Locasto. Cybersecurity education and assessment in edurange. *IEEE Security & Privacy*, (3):90–95, 2017.

[96] J. Werther, M. Zhivich, T. Leek, and N. Zeldovich. Experiences in cyber security education: The mit lincoln laboratory capture-the-flag exercise. In *CSET*, 2011.

[97] N. Wilhelmson and T. Svensson. *Handbook for planning, running and evaluating information technology and cyber security exercises*. Försvarshögskolan (FHS), 2011.

[98] S. Yari, S. Mäses, and O. Maennel. A method for teaching open source intelligence (OSINT) using personalised cloud-based exercises. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security: ICCWS 2020*. Academic Conferences and publishing limited, 2020.

Acknowledgements

What is the meaning of life? Once upon a time, this thought-provoking question was planned as the opening for this thesis. Thanks to my three supervisors, the final version of this thesis, and especially the introduction part, is much more concise and to the point.

So, without further ado, I'd like to thank, first and foremost, the German pilot with a Scandinavian name who has been my main supervisor. Thank you, Olaf, for numerous fruitful conversations and relentless constructive feedback, always aiming for the world-class quality (more data, more analysis, more explanation, more validation, more clarity, more references, and less meaningless filler words that actually do not add anything especially important to the already existing discussion that already examines various considerations regarding contrasting perspectives from already diversified aspects and frames of reference)!

Also, I'd like to thank my co-supervisors—Liina and Stefan—who have given their fair share of effort to guide my research to the current state. Having several perspectives from different supervisors really helped to consider different aspects of this interdisciplinary research. Memorable quote from Liina: *"The best PhD thesis is a finished one"*. Memorable quotes from Stefan: *"Sorry, but in social science words matter..." and "You can always use the IMPROVE button (meaning the button Delete)."* The wiser I get, the more I see how meaningful and true those quotes are.

Also, I am grateful for all of my colleagues—for those who have helped me and for those who have at least not obstructed my steps towards this academic milestone. The most significant helpers are already mentioned as my co-authors—and thank you all once again. Thanks for Rain for leading the cybersecurity group in our university and for creating a productive, but enjoyable working atmosphere.

Special thanks go to Adrian, thanks to whom numerous sentences have lost their syntax errors and obtained an extra touch of British charm. Also, several commas were relocated, and a cohort of (unnecessary) parentheses (and dashes) were obliterated—again thanks to him.

I strongly believe that humans learn best from positive examples. Thanks to Bernhards, Regina, Kadri, and other authors of recent PhD theses in the School of Information Technologies for giving me a better idea about how a finished PhD thesis should look like.

I thank the multitude of students in my courses who have taught me a lot. I also thank the substantial number of teachers who have inspired me to either be more like them or the opposite. In case I have missed someone who also contributed to this thesis or my well-being during its writing process, then this is done on purpose, similarly to all of my purposefully done mistakes in my lectures, just to make sure that you pay attention. Of course, I actually remember and I am grateful for your help. Also, I'd like to thank you. Yes, you, the reader who has stumbled upon this document. From all of this research, what was the most useful thing for you?

Last, but not least, I would like to thank Archimedeses—both the inspiring Greek inventor Archimedes who was an outstanding scientist centuries before the term "science" was established, and the organisation of the same name that has been distributing the resources of European Social Fund in Estonia. Their generous support (of the mentioned organisation, not the dead Greek) has allowed me to participate in several international conferences to promote the ideas presented in this thesis among the scientific community.

Abstract
Evaluating Cybersecurity-Related Competences through Simulation Exercises

There is a growing demand for effective cybersecurity education because creating and using various technologies in a safe and secure manner is becoming increasingly challenging. As it is difficult to improve what you cannot measure, having a proper competency evaluation is essential for effective cybersecurity education. Virtual cybersecurity simulations provide a scalable way to educate and evaluate people automatically.

Although virtual cybersecurity simulations are frequently used in education and training, it is often done without proper evaluation procedures. While simulation results might contribute towards a grade in university, they are rarely measured well enough to provide meaningful feedback. Previous research regarding cybersecurity simulation exercises has focused on solving related technical issues. There is a gap of knowledge regarding the use of the simulation exercises for measurable educational progress. The processes of evaluating competencies must be established before it is possible to measure the progress of the learners.

Case study research methods were used in the current work as they are suitable to a situation where there is not enough theoretical ground for defining specific hypotheses. This thesis establishes the processes for designing competency-driven cybersecurity simulation exercises where the results are mapped to a competency framework such as NIST NICE. Established individual competency profiles can contribute to early detection and intervention by targeting insufficient competency areas and identifying or building upon existing strengths to facilitate specialisation.

All the constructed case studies deal with designing some sort of method together with a proof of concept software solution. Design science research deals with the creation of successful proof of concept systems. The design science research methodology used in this thesis serves as a supporting framework for research and also for its structured presentation. The competency-driven exercise design approach suggested in this thesis is put to practice by several specifically created virtual simulations. In addition to illustrating the feasibility of competency-driven exercise design, these virtual simulations introduce novel technical solutions that demonstrate the versatility of serious games as an educational medium. The structured way of automatically evaluating competencies in virtual simulation exercises provides a scalable way to build a more effective, competency-based cybersecurity education.

Kokkuvõte
Küberturbe-alaste kompetentside hindamine simulatsiooniharjutuste abil

Infoühiskonna kasvav sõltuvus erinevatest andmeid töötlevatest seadmetest on tekitanud suureneva vajaduse küberturbealaste kompetentside järele, mille abil neid seadmeid turvaliselt luua ja kasutada. Vajalike küberturbe kompetentsidega inimeste harimiseks on aga tähtis aru saada, kuidas neid kompetentse paremini mõõta. Ilma regulaarsete mõõtmisteta on keeruline leida tõhusaid viise, kuidas küberturbealaseid kompetentse õpetada.

Antud töö eesmärgiks oli kirjeldada, kuidas küberturbealaseid kompetentse simulatsiooniharjutuste abil paremini hinnata. Simulatsiooniharjutuste all mõeldakse antud töös virtuaalseid simulatsioone, mida sooritatakse sageli virtuaallaboritena. Virtuaallabor on pilvekeskkonnas asuv kogum virtuaalmasinaid, millele õppijal on kaugjuurdepääs.

Erinevaid virtuaalseid simulatsioone kasutatakse küberturbe õpetamisel sageli, sest need võimaldavad õppijatel omandada vajalikke praktilisi oskusi. Paraku on küberturbe simulatsioonid harva seotud reaalse õppeprotsessi tulemustega. Küberturbe hariduse parendamiseks tuleks kasuks erinevate kompetentside täpsem ja süstemaatilisem hindamine. Töö eesmärgi saavutamiseks tõstatati järgmised uurimisküsimused:

- Kuidas küberturbealaseid kompetentse liigitada/kategoriseerida?
- Milliseid oskuste hindamiseks relevantsed meetrikaid on võimalik simulatsiooniharjutuste abil koguda?
- Milliseid küberturbealaseid oskuseid on võimalik simulatsiooniharjutuste abil hinnata?
- Kuidas luua küberturbealaste oskuste mõõtmiseks sobilikke simulatsiooniharjutusi?

Uurimisküsimustele vastuste leidmiseks rakendati kvalitatiivset juhtumianalüüsi ja disainiteaduse uurimismetoodikat. Küberturbe valdkonna süsteemsemaks käsitlemiseks loodi kontseptuaalne *Cybersec-Tech* akna mudel (kompetentside maatriks), mis aitab kohaseid kompetentse paremini liigitada. *Cybersec-Tech* mudel jagab kompetentsid nelja kategooriasse vastavalt sellele, kuivõrd küberturbe-spetsiifilised ja tehnilised need on. See võimaldab erineva taustaga inimestel küberturbe harjutuste loomist paremini koordineerida, sest tekib selgem arusaam loodava harjutuse ulatusest ja fookusest.

Harjutuse üldise eesmärgi seadmisest jääb aga väheks, kui ei õnnestu harjutuse käigus kasutatud oskuseid täpsemalt mõõta. Sageli piirdutakse küberturbe harjutuste hindamisel osalejate tagasisidega, mis ei anna simulatsiooni jooksul demonstreeritud oskuste kohta objektiivset hinnangut. Oskuste paremaks hindamiseks on vaja harjutuse käigus mõõdetud tegevuste tulemused ühendada süstemaatilise oskuste kogumiga. Raamistikud, nagu näiteks enamkasutatud NIST NICE, pakuvad kogumit erinevatest kompetentsidest, mis on seotud küberturbealaste töörollidega. NIST NICE raamistikku aluseks võttes kirjeldati küberturbe harjutuste loomise protsessid võimaldamaks luua simulatsioone, mis on otseselt seotud kindlate küberturbe töökirjeldustega.

Automaatselt hinnatavad virtuaallaborid võimaldavad küberturbe haridust skaleerida, sest rohkemate inimeste õpetamiseks on vaja peamiselt suurendada masinressurssi. Töökirjeldustega seotud tulemused võimaldavad simulatsioonides osalejatel hinnata enda sobivust erinevatele küberturbealastele töökohtadele. Õppematerjalide sidumine reaalsete töökohas vajaminevate kompetentsidega aitab mõtestada kogu õppeprotsessi, sest õppija näeb paremini, milleks omandatavaid kompetentse võib tulevikus vaja minna.

Lisaks kompetentispõhiste küberturbealaste harjutuste loomisprotsesside kirjeldamisele, loodi antud doktoritöö käigus ka mitmeid uuenduslikke virtuaallaboreid, mis näitasid, et virtuaalsed simulatsioonid ei pea tingimata olema isolatsioonis toimuvad tehnilised harjutused nagu neid siiani sageli on kasutatud. Virtuaalsed simulatsioonid võivad olla ühenduses avalike teenustega Internetis ja mõõta ka mitte-tehnilisi kompetentse.

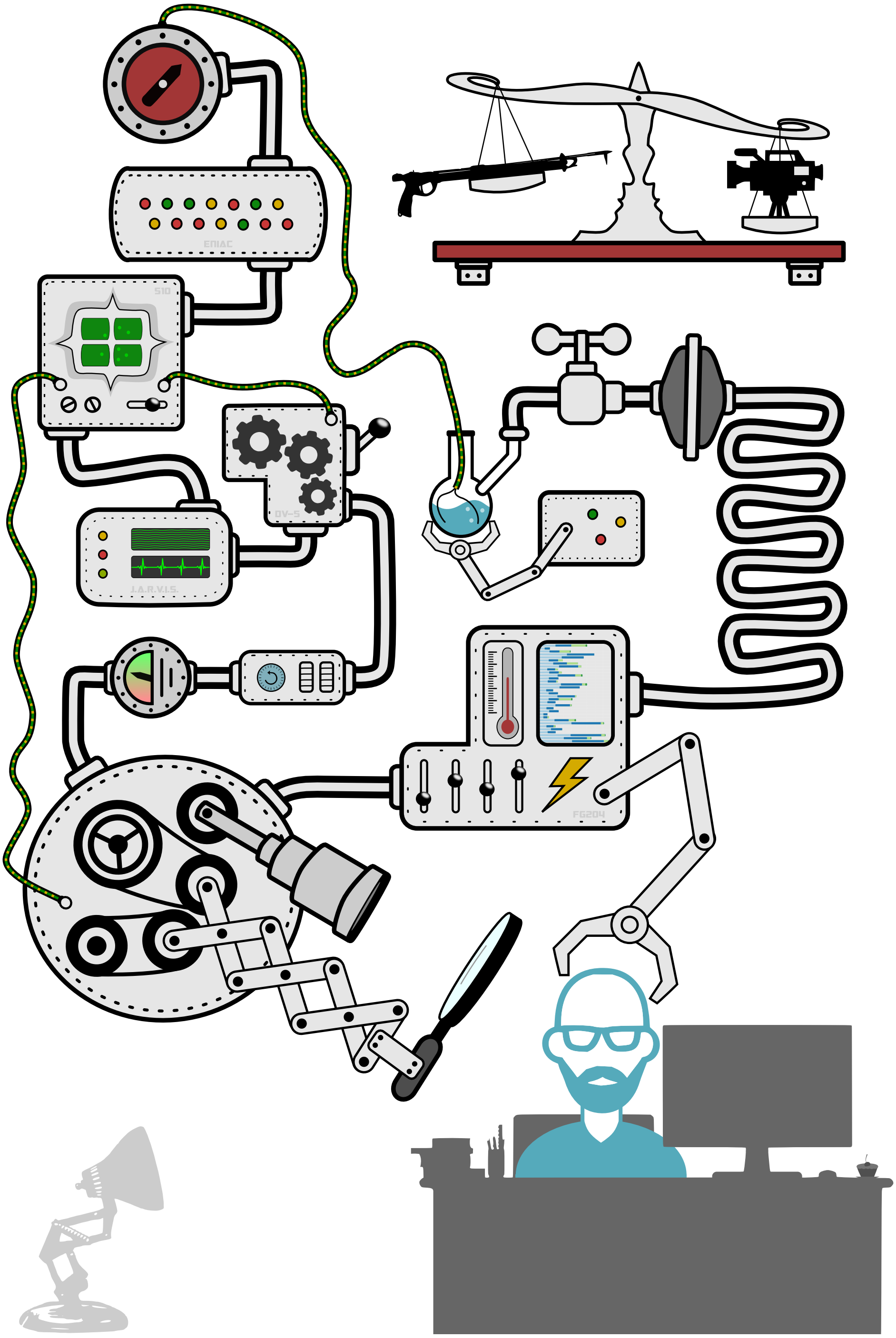
Tehnilise ja küberturbe-spetsiifilise teema näidislabor keskendus OSINT (*Open Source Intelligence* ehk leheluure) oskustele. Kuna OSINT on suuresti sõltuv avalikult kättesaadavast infost, on selleteemalisi harjutusi üldjuhul keeruline tõepäraselt simuleerida. Realistliku simulatsiooni loomiseks kasutati vahendusrünnet, mis võimaldas spetsiifilistele labori kaudu vaadeldavatele avalikele veebilehtede sisu muuta. Nii sai lisada avalikele veebilehtedele ainult laborikeskkonnast nähtavaid infokilde (antud juhul 16-märgiline tekst), mille avastamine näitas osaleja teatud oskuseid. Kuna infokillu täpne sisu oli igal laboris osalejal erinev, vähendas see võimalusi teise osaleja leidu jäljendada ise ülesannet tegemata. Tulemusena loodi meetod skaleeruva tehnilise harjutuse loomiseks, mis suudab edukalt kasutada laborikeskkonnast väljaspool asuvaid süsteeme, kuid sellegipoolest muuta harjutused piisavalt individuaalseks, et leitud vastuste jagamine ei oleks triviaalne.

Vähem tehnilise ja vähem küberturbe-spetsiifilise teema näidiseks loodi meetod kübereetilise käitumise hindamiseks virtuaallaborite abil. Tehnilistele virtuaallaboritele lisati eetiline dilemma (vajaliku loa küsimine) ja mõõdeti tudengite käitumist. Meetod näitas, kuidas on võimalik tehnilistele virtuaallaboritele lisada ka mitte-tehniliste kompetentside mõõtmise aspekte. Kuna lisaks küberturbe tehnilistele kompetentsidele on aina enam hakatud väärtustama ka üldisi mitte-tehnilisi kompetentse, siis oleks loogiline haridussüsteemis ka selliseid kompetentse arendada ja mõõta. Loodud meetod näitas, et virtuaalsetele simulatsioonidele on võimalik lisada ka mitte-tehniliste aspektide mõõtmine. Loodetavasti inspireerib see lähenemine tulevikus rohkem tähelepanu pöörama ka mitte-tehniliste kompetentside mõõdetavale arendamisele.

Erinevad kompetentsipõhised simulatsioonid leidsid rakendust sissejuhatava küberturbe kursuse raames, kus koguti tudengite tulemused küberturbe virtuaallaboritest, testidest ja kodutöödest. Erinevad sooritused ühendati süsteemselt NIST NICE kompetentside raamistikuga. See võimaldas luua individuaalsed kompetentsikaardid, mis näitavad tudengite teadmisi ja oskusi kindlates küberturbealaste töörollidega seotud valdkondades. Tehtud töö näitab, kuidas on võimalik automaatselt hinnatavate kompetentsipõhiste virtuaallaborite kasutamine õppetöös üheskoos teiste testide ja kodutöödega.

Käesolev töö on loonud hea aluse laialdasemaks kompetentsipõhise õppe rakendamiseks küberturbe hariduses. Doktoritöö põhineb seitsmel akadeemilisel artiklil, mis käsitlevad erinevaid uuenduslikke meetodeid ja protsesse kompetentsipõhiste küberturbe harjutuste loomiseks. Loodetavasti aitavad loodud protsessimudelid ja uued teadmised laborite loomise ja kasutamise kohta küberturbe haridust edendada. Nii oleks maailmas rohkem kompetentseid inimesi, kes oskaks hoida infoühiskonda piisavalt turvalisena.

Graphical abstract



Appendix 1

I

S. Mäses, B. Hallaq, and O. Maennel. Obtaining better metrics for complex serious games within virtualised simulation environments. In *ECGBL 2017 11th European Conference on Game-Based Learning*, pages 428–434. Academic Conferences and publishing limited, 2017

Obtaining Better Metrics for Complex Serious Games Within Virtualised Simulation Environments

Sten Mäses[1], Bil Hallaq[2], Olaf Maennel[1]
[1]Department of Software Science, Tallinn University of Technology, Estonia
[2]Cyber Security Department, WMG - University of Warwick, U.K.
sten.mases@ttu.ee
bh@warwick.ac.uk
olaf.maennel@ttu.ee

Abstract: Recent technological advancements are providing significant results in enriching learning through serious games in the field of information and communication technology (ICT). One such advancement is the ability to create highly complex virtualised environments that realistically simulate organisations' ICT systems, enabling participants to develop practical hands-on skills in a controlled environment. During such exercises, a critical component is the ability to track individual participants progress. This requires an evaluation system to be in place. Within traditional computer gaming environments, it is relatively easy to automate the tracking and capture of a specific player's moves, clicks and other interaction. Within simulations of serious games such as those used for training defence and attack mitigation techniques in a computer network, tracking such activities in an automated manner is significantly more complex. Using serious games in the field of cybersecurity as an example, we provide in this work a mapping of different types of metrics for serious games run within virtual lab environments and suggest various ways in which they can be measured. This work will assist serious game designers, developers, organisers and assessors obtain a greater understanding of the state of the art possibilities for measuring the performance of participants. It will also enable researchers to build a solid foundation in which they can develop new approaches for more efficient learning through virtualized simulations.

Keywords: serious games analytics, game-based learning, cybersecurity, computer simulation environments

1. Introduction

Cybersecurity is one of the fields that faces an ongoing global resource crisis (Loeb, 2015). Due to the lack of qualified personnel in this rapidly growing field, it is increasing difficult to find technically proficient individuals (Evans, K. & Reeder, 2010). Like in other fields, a qualified cybersecurity specialist needs sufficient knowledge together with practical skills (Assante and Tobey, 2011). Knowledge acts as an enabler for different skills, and skills can be directly applied to solve everyday tasks and problems. Obtaining the relevant skills though, especially in complex areas such as cybersecurity, is often not a simple task to be achieved. As with many fields, it is risky to let inexperienced people access systems in order to learn (Lateef, 2010). Therefore, various educational simulations have been developed to address this issue.

Games such as chess and Go have been used for centuries to train strategical thinking, but the skills learnt from these games usually cannot be transferred straight to the real world. For example, while experience of chess can be inspirational for configuring a firewall, it is surely not sufficient. The big question with simulations is always, up to what extent are the skills learnt from the simulation transferrable to a real-life situation? In that perspective, the field of computer science has a significant advantage. Due to the popularity of cloud based systems, most technical skills needed from an IT specialist are connected to virtualised environments. Therefore, it is possible to create virtualised simulation environments that are almost identical to, or in some scenarios identical to real-life systems.

The term serious games is often used in order to describe games with a non-entertaining purpose (Djaouti, Alvarez and Jessel, 2011). In this work, we take the cybersecurity field as an example and show how serious games can help us develop and measure relevant skills. For skill classification, we use Bloom's taxonomy (Starr *et al.*, 2008).

2. Related work

The concept of using digital games to support learning is not new. For more than a decade, many researchers have published numerous essays, articles and books on the topic of digital game-based learning (Eck, 2006). This research is largely motivated by the need to create more engaging learning environments. The educational effects of many simulations have also been researched quite widely throughout diverse fields such as flight simulators (Hays *et al.*, 1992), nursing (Jacobs, Beyer and Carter, 2017), and cybersecurity (Furfaro *et al.*, 2017).

There are also multiple serious games in the cybersecurity field that are used for the training of specialists. Tabletop exercises can be used to demonstrate the interrelationships of the technical, procedural and human aspects of cybersecurity (Ottis, 2014). In addition to these there are numerous complex technical exercises designed to train security experts who protect critical IT systems such as Locked Shields (NATO CCDCOE, 2017), CYRAN (Hallaq *et al.*, 2016), U.S. Service Academies Cyber Defence Exercise (Carlin and Manson, 2011), and Cyber Shield (Henshel *et al.*, 2016).

In addition to the complex large-scale technical exercises there has been an increasing number of smaller scale exercises using virtualised simulation environments for different educational goals. Some are more general (Cano *et al.*, 2016) and others more specific, e.g. a lab for learning the concepts connected to denial of service (Ledford, Mountrouidou and Li, 2016) or assessment of cybersecurity issues in Internet of Things (IoT) scenarios (Furfaro *et al.*, 2017). There are also both commercial (Buttyán, Félegyházi and Pék, 2016) and open source platforms (Ernits, Tammekänd and Maennel, 2015) for managing virtualised simulation environments.

Serious games which are run within virtualised simulation environments appear to be the best solution to ensure resilient, efficient and scalable cybersecurity exercises. They consist of different virtual machines (VM-s) that are interconnected to one or many networks. Similar technology is being used in large scale production environments for offering a diverse selection of cloud-based services – therefore the skills learnt from VM-based lab environments are highly applicable also outside of the lab.

Within this section we have reviewed various simulation environments and the complexities associated with them. While many projects are benefitting from the advantages of VM-based simulation environments, there is a lack of systematic approach to identify general guidelines for performance measurement of the participants of such exercises. Much of the work in this area typically happens in isolation and those that are federated or openly available, such as DETERLab (Wroclawski *et al.*, 2016), are not more widely adopted.

On the other hand, there is a growing body of literature on the topic of Serious Games Analytics (SEGA), but there is a shortage of research considering applying SEGA to VM-based simulation environments. This paper aims to clarify and structure some core concepts regarding SEGA in VM-based simulation environments. Within the next section we identify the types of virtual environments available and the benefits and drawbacks of each.

3. VM-based environments

VM-based environments can be divided based on the user's access level to the system.

In case of restricted access, the user has only predefined operations to choose from. This way the VM-based environment can stay on the background and provide the user with a simple user interface that is convenient for people with lower technical skills. While it can be a good starting point for a deeper technical understanding, it is more focused on the first two levels of Bloom's taxonomy (Starr *et al.*, 2008), recall and comprehension.

In case of full access to the virtual machine, the user can choose their own methods and tools for tackling the task. This option is very realistic; therefore, it can be used for learning skills from all levels of Bloom's taxonomy. At the same time, it can be overwhelming for users with lower technical skills. Also, it demands relatively high-throughput bandwidth in the case of doing the simulation remotely.

There are also options with selected access where the user can write their own commands through the browser interface, but the full graphical interface of the virtual machine is hidden. This approach can be considered as a compromise. All skill levels of the Bloom's taxonomy can still be reached, but the simulation is less realistic than in the case of full access to the virtual machine, because the user interface is likely slightly modified based on the simulation environment (e.g. browser constraints).

4. Performance metrics

Serious games are designed to support knowledge acquisition and/or skill development (Loh and Yanyan Sheng, 2013). Therefore, it is essential to assess the performance of the players. The performance is usually characterised by a score, that is comparable between the participants of the same serious game (Loh, Sheng and Ifenthaler, 2015). The input for the score comes from different measurement points. As different measurement points often measure similar metrics, then it is reasonable to form categories that encompass similar measurement points as illustrated in the **Figure 1**. To calculate the total score S_{total}, the input from different measurement points can be weighed as follows:

$$S_{total} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where w_i is the weight and x_i is the performance metric value from one measurement point. Similarly, different weights can be assigned to different categories to prioritise some of them over the others.

As an example, let us imagine a cybersecurity-related serious game with the following two performance metric categories: availability of services and incident reporting. If the goal of this serious game is to promote the importance of proper incident reporting, then the weight of the according category can be several times higher than the weight of the other categories (in the current example, the availability of services). That way, the participants are encouraged to focus more on incident reporting and less on the availability of services. Of course, this kind of prioritisation only works if the participants are at least to some extent aware of the scoring system, not in case of a stealth assessment (DeRosier, Craig and Sanchez, 2012).

The weight for any performance metric is highly dependable on the game and should be fine-tuned during the testing phase based on the real results.

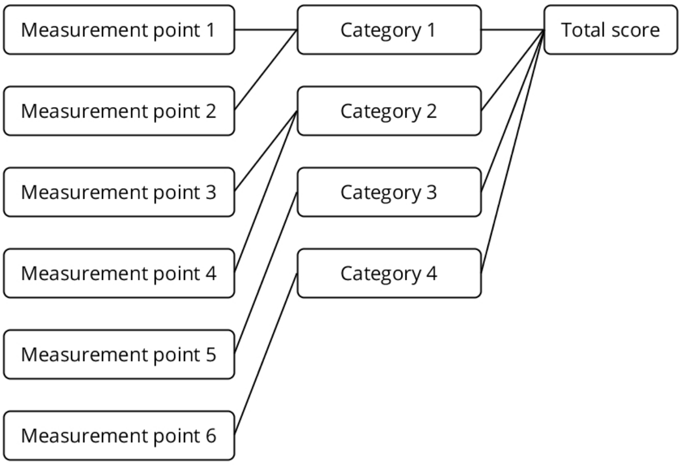


Figure 1: *Categorising performance metrics*

The total score is necessary in the case where the VM-based serious game is used for a competition or formal assessment and a final ranking list is required. A further analysis by performance metric categories (and potentially sub-categories) can provide much more valuable insights about the strengths and weaknesses of the participant (or participating team).

Each measurement point contributes to the final scoring of the serious game by measuring a specific metric.

Some metrics are measuring whether the main objective of a task was completed, while others give additional information about the way the participant achieved (or did not achieve) the required objective. The completion of a task can be measured in a binary (pass/fail) or continuous scale. Any continuous scale measurement can be turned into a binary measurement by introducing a threshold (e.g. defining the border between fail and pass).

Those metrics can be measured in various ways. In the following list, there are some of the commonly used metrics to consider.

4.1 Direct input

The objective of a VM-based serious game participant can be to find some specific information using the virtual machines and then report this information to be evaluated. Reporting can be done through a side-channel, such as a Moodle platform[1] that the user can access outside of the virtual machines. For example, an ICT security project DECAMP in Munich (Alexandru Soceanu, Maksym Vasylenko and Alexandru Gradinaru, 2017) developed a lab based on Moodle education platform interfaced with OpenStack[2] cloud computing platform.

Using the metrics based on direct input is more reliable if the objectives are slightly customised for each participant. For example, the task for each participant can be to decrypt a message, but the content of the message to be submitted for evaluation could be different for each participant. This way, it can be confirmed, that no participant reached the result by simply getting the required information from others.

4.2 Automated scoring script

Checking the objectives of the VM-based serious game can be automated. For example, an automated script can evaluate the uptime of a service or check the contents of a specific configuration file. When using automated scoring, it is important to limit the participant's access to the scoring script. For example, if the scoring script resides in the same virtual machine which the user has full access to, then it is not difficult to alter the scoring script. Therefore, it would be possible for a person to get maximum points without even completing the actual given objective.

To protect the integrity of the scoring system, it is possible to place the scoring script into another virtual machine (that the user has no access to), or to protect the scoring script by limiting the user administrative rights.

4.3 Time

Time spent on VM-based serious game objectives can give valuable insights about the participant's abilities. Average time spent on every task can give a good basis for further analysis. In addition, it can be used as a trigger for displaying hints to participants who have gotten stuck in a pre-defined scenario.

It should be considered though, that the time spent on different tasks can be influenced by many external factors, e.g. user's hardware, Internet bandwidth, coffee breaks etc.

Also, while usually a faster task completion indicates a more skilled participant, it should be considered that unrealistically fast times might indicate a fault in the system or a malicious participant instead.

4.4 String similarity metrics

Comparing unknown performance against known experts can be done using string similarity metrics such as Levenshtein distance (Loh and Sheng, 2015). String similarity metrics can be used to compare activity logs, e.g. history of typed commands by the participant. This allows for evaluation of the efficiency of the participant's actions.

4.5 Tools

In VM-based simulation, the user often has full control over the machine and therefore also freedom to choose amongst multiple tools. The choice of a specific software (e.g. Nmap[3]) or a programming language (e.g. Bash or Python) can give extra information about the participant's skillset.

[1] Moodle, https://moodle.com

[2] OpenStack Project, http://www.openstack.org

[3] Nmap Security Scanner, https://nmap.org

5. Mapping of skills

Measurements are meaningless unless they are connected to specific skills. Therefore, it is necessary to define the particular skillset that the serious game should develop and measure. NICE Cybersecurity Workforce Framework (Newhouse *et al.*, 2016) could be used as a reference for mapping relevant cybersecurity skills. For ICT occupations in general, the e-Skills Match framework (Fernández-Sanz, Gómez-Pérez and Castillo-Martínez, 2017) can be used for integrating the existing ICT related reference schemes and standards. Using widely established standards and frameworks as a basis, helps to compare the performance metrics across different systems and organisations.

Skill differentiation is also important. Technological advances enable us to keep track of more data and that enables more granular tracking of performance and abilities (Umbleja *et al.*, 2014). With the help of automated scoring systems, it is possible to give individual and highly specific feedback for each participant of a virtual simulation. **Figure 2** shows how skills can be linked to different performance measurement categories. Additionally, different granular skills can be grouped together as skill sets. This provides a high-level overview of a participant's skills and can be easily communicated. For example, it would be confusing for a job ad to specify hundreds of granular skills that are required for a position. Instead a wider skill set is often described. If the same skill set is defined based on a VM-based serious game skills and linked to measurements, then it is possible to have an automated and scalable system that can quickly find the job applicant with the most fitting technical skill sets.

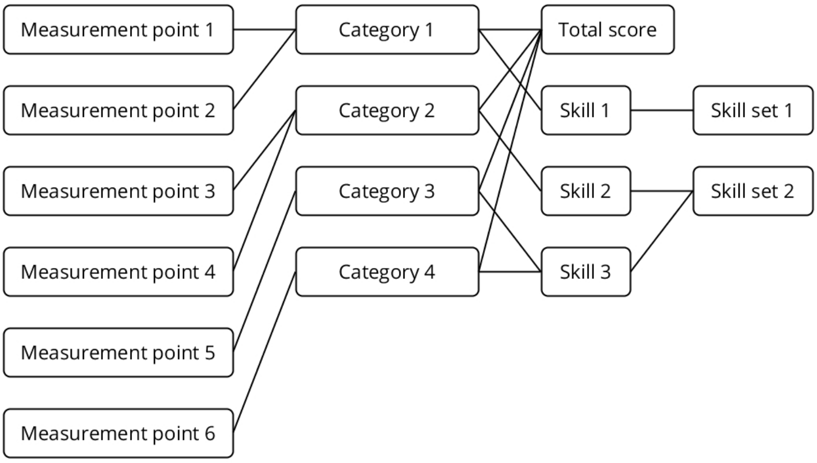


Figure 2: *Categorised performance metrics linked to skills*

In the current example (**Figure 2**), the total score is not affected by the skill sets. The scores for each skill set help to give participants detailed and individual feedback. In a different context, a total score can be also based on the similarity to the pre-defined skill sets (e.g. when finding the job applicant with best fitting technical background). It is important to note that there can be various types of connections between measurement points, categories, skills and skill sets. For example, there can be many measurement points giving input to a common category. Similarly, many categories of performance metrics can be connected to one skill (or one skill to many categories). **Figure 3** provides an example for using the described framework for mapping the performance metrics and relevant skills.

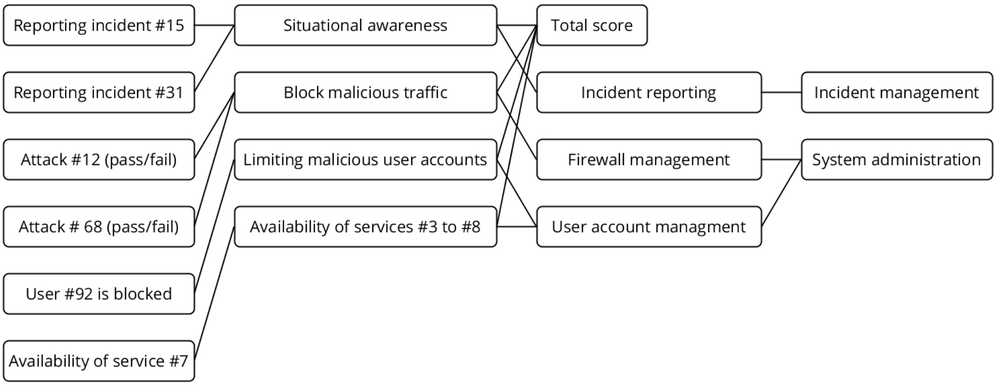


Figure 3: *Sample of categorised performance metrics linked to skills*

Figure 3 illustrates a mapping of performance metrics and skills in a hypothetical scenario. All the numbers (e.g. service #7) are arbitrary and for illustrative purposes only. There are six measurement points. Out of these, two measurement points are focusing on incident reporting. Those two give the input for the situational awareness category scoring. Optionally, the weight of reporting incident #15 and #31 could vary. For example, reporting incident #15 could be considered two times more important than reporting incident #31 and the weights could be set accordingly. Situational awareness together with the other performance metric categories gives input to the total score. In addition to being used for calculating the total score, it can be seen how situational awareness is used as an indication for the incident reporting skill. In the current example, the incident reporting skill is the only member of the incident management skill set. In reality, there can be numerous other skills that form together an input to the incident management skill set.

6. Future research

In this work, we defined metrics that are mostly based on common input sources such as the keyboard or the mouse. In the future, more inputs could be considered. Video feed from the in-built camera can be used to conduct eye tracking (Galdi *et al.*, 2016). Audio from the microphone can be analysed to measure user's stress level, group dynamics and emotional tonalities of the communication by voice (Shiva Prasad, Kodanda Ramaiah and Manjunatha, 2017). Keyboard input can be analysed further for purposes such as verifying the user identity (Bhattasali *et al.*, 2016).

Additionally, more research should be done regarding the dynamic analysis of serious game data. Cognitive principles for effective learning include creating learning scenarios that keep the participants in a zone between too easy and too difficult (Greitzer, Kuchar and Huston, 2007). Automatic analysis of performance metrics can allow the computerised serious game to adapt to individual needs of each participant and therefore provide a more engaging experience.

7. Conclusions

Serious games analytics (SEGA) is a growing field with diverse applicable implementations. We have identified computer science and cybersecurity as fitting areas to be linked with serious games analytics due to their extremely life-like VM-based simulation environments. To address the lack of systematic approach for connecting SEGA with VM-based serious games, we have outlined the main relevant performance metrics for measuring the skills of participants of VM-based serious games. The key measurement points are brought out together with their strengths and pitfalls. Furthermore, the approach for connecting between performance metric categories and skills is described.

In this paper, we have used cybersecurity related serious games as an example, but the same principles can be carried over to multiple other fields, especially to those that are connected to computer science.

8. Bibliography

Alexandru Soceanu, Maksym Vasylenko and Alexandru Gradinaru (2017) 'Improving Cybersecurity Skills Using

Network Security Virtual Labs', in *Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 Vol II, IMECS*. Hong Kong. Available at: http://www.iaeng.org/publication/IMECS2017/IMECS2017_pp594-599.pdf (Accessed: 3 May 2017).

Assante, M. J. and Tobey, D. H. (2011) 'Enhancing the Cybersecurity Workforce', *IT Professional*, 13(1), pp. 12–15. doi: 10.1109/MITP.2011.6.

Bhattasali, T., Panasiuk, P., Saeed, K., Chaki, N. and Chaki, R. (2016) 'Modular logic of authentication using dynamic keystroke pattern analysis', *AIP Conference Proceedings*, 180012(101). doi: 10.1063/1.4951959.

Buttyán, L., Félegyházi, M. and Pék, G. (2016) 'Mentoring talent in IT security – A case study', in *USENIX Workshop on Advances in Security Education (ASE '16)*. Available at: https://www.usenix.org/system/files/conference/ase16/ase16-paper-buttyan.pdf (Accessed: 2 May 2017).

Cano, J., Hernandez, R., Ros, S. and Tobarra, L. (2016) 'A distributed laboratory architecture for game based learning in cybersecurity and critical infrastructures', in *2016 13th International Conference on Remote Engineering and Virtual Instrumentation (REV)*. IEEE, pp. 183–185. doi: 10.1109/REV.2016.7444461.

Carlin, A. and Manson, D. (2011) 'A League of Our Own : The Future of Cyber Defense Competitions', *Communications of the IIMA*. International Information Management Association, 11(2), pp. 1–11. Available at: http://scholarworks.lib.csusb.edu/ciima/vol11/iss2/1 (Accessed: 3 May 2017).

DeRosier, M. E., Craig, A. B. and Sanchez, R. P. (2012) 'Zoo U : A Stealth Approach to Social Skills Assessment in Schools', *Advances in Human-Computer Interaction*. Hindawi Publishing Corp., 2012, pp. 1–7. doi: 10.1155/2012/654791.

Djaouti, D., Alvarez, J. and Jessel, J.-P. (2011) 'Classifying serious games: The G/P/S model', *Handbook of research on improving learning and motivation through educational games: Multidisciplinary approaches*. IGI Global, (2005), pp. 118–136. doi: 10.4018/978-1-60960-495-0.ch006.

Eck, R. Van (2006) 'Digital Game-Based Learning : It â€™ s Not Just the Digital Natives Who Are Restless ….', *Educause Review*. ACM, 41(2), pp. 1–16. doi: 10.1145/950566.950596.

Ernits, M., Tammekänd, J. and Maennel, O. (2015) 'i-tee: A fully automated Cyber Defense Competition for Students', *ACM SIGCOMM Computer Communication Review*. ACM, 45(5), pp. 113–114. doi: 10.1145/2829988.2790033.

Evans, K. & Reeder, R. (2010) *A Human Capital Crisis in Cybersecurity: Technical Proficiency Matters - A Report of the CSIS Commission on Cybersecurity for the 44th Presidency*. Available at: www.csis.org (Accessed: 2 May 2017).

Fernández-Sanz, L., Gómez-Pérez, J. and Castillo-Martínez, A. (2017) 'e-Skills Match: A framework for mapping and integrating the main skills, knowledge and competence standards and models for ICT occupations', *Computer Standards & Interfaces*, 51, pp. 30–42. doi: 10.1016/j.csi.2016.11.004.

Furfaro, A., Argento, L., Parise, A. and Piccolo, A. (2017) 'Using virtual environments for the assessment of cybersecurity issues in IoT scenarios', *Simulation Modelling Practice and Theory*, 73, pp. 43–54. doi: 10.1016/j.simpat.2016.09.007.

Galdi, C., Nappi, M., Riccio, D. and Wechsler, H. (2016) 'Eye movement analysis for human authentication: a critical survey', *Pattern Recognition Letters*, 84, pp. 272–283. doi: 10.1016/j.patrec.2016.11.002.

Greitzer, F. L., Kuchar, O. A. and Huston, K. (2007) 'Cognitive science implications for enhancing training effectiveness in a serious gaming context', *Journal on Educational Resources in Computing*. ACM, 7(3), p. 2–es. doi: 10.1145/1281320.1281322.

Hallaq, B., Nicholson, A., Smith, R., Maglaras, L., Janicke, H. and Jones, K. (2016) 'CYRAN: A Hybrid Cyber Range for Testing', in *Security Solutions and Applied Cryptography in Smart Grid Communications*. IGI Global, pp. 226–241. doi: 10.4018/978-1-5225-1829-7.ch012.

Hays, R. T., Jacobs, J. W., Prince, C. and Salas, E. (1992) 'Flight simulator training effectiveness: A meta-analysis.', *Military Psychology*. Lawrence Erlbaum, 4(2), pp. 63–74. doi: 10.1207/s15327876mp0402_1.

Henshel, D. S., Deckard, G. M., Lufkin, B., Buchler, N., Hoffman, B., Rajivan, P. and Collman, S. (2016) 'Predicting proficiency in cyber defense team exercises', in *MILCOM 2016 - 2016 IEEE Military Communications Conference*. IEEE, pp. 776–781. doi: 10.1109/MILCOM.2016.7795423.

Jacobs, R., Beyer, E. and Carter, K. (2017) *Interprofessional simulation education designed to teach occupational therapy and nursing students complex patient transfers*, *Journal of Interprofessional Education & Practice*. doi: 10.1016/j.xjep.2016.12.002.

Lateef, F. (2010) 'Simulation-based learning: Just like the real thing.', *Journal of emergencies, trauma, and shock*. Medknow Publications and Media Pvt. Ltd., 3(4), pp. 348–52. doi: 10.4103/0974-2700.70743.

Ledford, H., Mountrouidou, X. and Li, X. (2016) 'Denial of Service Lab for Experiential Cybersecurity Learning in Primarily Undergraduate Institutions', *Journal of Computing Sciences in Colleges*, 32(2), pp. 158–164. Available at: http://delivery.acm.org/10.1145/3020000/3015088/p158-

ledford.pdf?ip=193.40.244.196&id=3015088&acc=PUBLIC&key=D2103A8F5527A3D9.5764A7F6B87355B6.4D4702B0C3E38B35.4D4702B0C3E38B35&CFID=757649657&CFTOKEN=70574058&__acm__=1493719799_5ff096bd9d26ea17fee04 (Accessed: 2 May 2017).

Loeb, M. (2015) 'Cybersecurity talent: Worse than a skills shortage, it's a critical gap', *The Hill*, 17 April. Available at: http://thehill.com/blogs/congress-blog/technology/239113-cybersecurity-talent-worse-than-a-skills-shortage-its-a.

Loh, C. S. and Sheng, Y. (2015) 'Measuring the (dis-)similarity between expert and novice behaviors as serious games analytics', *Education and Information Technologies*. Springer US, 20(1), pp. 5–19. doi: 10.1007/s10639-013-9263-y.

Loh, C. S., Sheng, Y. and Ifenthaler, D. (2015) 'Serious Games Analytics: Theoretical Framework', in *Serious Games Analytics*. Cham: Springer International Publishing, pp. 3–29. doi: 10.1007/978-3-319-05834-4_1.

Loh, C. S. and Yanyan Sheng (2013) 'Performance metrics for serious games: Will the (real) expert please step forward?', in *Proceedings of CGAMES'2013 USA*. IEEE, pp. 202–206. doi: 10.1109/CGames.2013.6632633.

NATO CCDCOE (2017) *Locked Shields 2017*. Available at: https://ccdcoe.org/locked-shields-2017.html (Accessed: 2 May 2017).

Newhouse, B., Keith, S., Scribner, B. and Witte, G. (2016) 'NICE Cybersecurity Workforce Framework (NCWF): National Initiative for Cybersecurity Eduction (NICE)', *Draft NIST Special Publication 800-181*. Available at: http://csrc.nist.gov/publications/drafts/800-181/sp800_181_draft.pdf (Accessed: 3 May 2017).

Ottis, R. (2014) 'Light Weight Tabletop Exercise for Cybersecurity Education', *Journal of Homeland Security and Emergency Management*, 11(4), pp. 579–592. doi: 10.1515/jhsem-2014-0031.

Shiva Prasad, K. M., Kodanda Ramaiah, G. N. and Manjunatha, M. B. (2017) 'Speech Features Extraction Techniques for Robust Emotional Speech Analysis/Recognition', *Indian Journal of Science and Technology*, 10(3). doi: 10.17485/ijst/2017/v10i3/110571.

Starr, C. W., Manaris, B., Stalvey, R. H., Starr, C. W., Manaris, B. and Stalvey, R. H. (2008) 'Bloom's taxonomy revisited', *ACM SIGCSE Bulletin*. ACM, 40(1), p. 261. doi: 10.1145/1352322.1352227.

Umbleja, K., Kukk, V., Jaanus, M. and Udal, A. (2014) 'New concepts of automatic answer evaluation in competence based learning', in *IEEE Global Engineering Education Conference, EDUCON*. IEEE, pp. 922–925. doi: 10.1109/EDUCON.2014.6826207.

Wroclawski, J., Benzel, T., Blythe, J., Faber, T., Hussain, A., Mirkovic, J. and Schwab, S. (2016) 'DETERLab and the DETER Project', in *The GENI Book*. Cham: Springer International Publishing, pp. 35–62. doi: 10.1007/978-3-319-33769-2_3.

Appendix 2

II

S. Mäses, L. Randmann, O. Maennel, and B. Lorenz. Stenmap: Framework for evaluating cybersecurity-related skills based on computer simulations. In *Learning and Collaboration Technologies. Learning and Teaching*, pages 1–13. Springer, 2018

Stenmap: Framework for Evaluating Cybersecurity-Related Skills Based on Computer Simulations

Sten Mäses[1 [0000-0002-1599-8649]], Liina Randmann[1],
Olaf Maennel[1 [0000-0002-9621-0787]], Birgy Lorenz[1 [0000-0002-9890-2111]],

[1] Tallinn University of Technology, Tallinn, Estonia
{sten.mases, liina.randmann, olaf.maennel, birgy.lorenz}@ttu.ee

Abstract. Cybersecurity exercises have become increasingly popular for training and assessing practical skills of information security. Nevertheless, the main focus of those exercises still tends to be on achieving completion or winning condition, not on understanding the individual skill set of each participant. This paper builds upon related work in creating and implementing cybersecurity exercises and proposes a revised process that includes competency mapping. A new conceptual framework called "Stenmap" is introduced that turns the results of cybersecurity simulations into a more meaningful evaluation of each participant's individual skills. Cybersec-Tech window is introduced as a tool for discussing the high-level classification of cybersecurity-related skills and reaching a common understanding among exercise organisers coming from diverse backgrounds. An Estonian national cybersecurity competition is used as an example for implementing the suggested process and Stenmap framework.

Keywords: Cybersecurity Simulations, Cybersecurity Exercises, Competency Mapping, Serious Games.

1 Introduction

Computer simulations are widely adopted by academia and computer security industry for augmenting information security education [1]. Realistic hands-on exercises enable to assess the practical skills of participants, contrary to many professional cybersecurity certification exams (such as CISSP, CISA and CISM) that are mostly based on knowledge tests [2]. Cybersecurity simulations, therefore, provide a useful tool for finding necessary cybersecurity specialists to fill in the growing staffing need in this area [3]. While there is quite a wide variety of different cybersecurity exercises, not much attention has been given to systematic skill assessment and improvement. A more standardised approach could enable a better comparability between the scores of different exercises, therefore facilitating the matching process of qualified specialists with suitable job positions.

This paper looks at previous work in the field of cybersecurity exercises and aims to unify different approaches into one revised process for creating and implementing cybersecurity exercises. The main goal of the paper is to provide a conceptual theoretical framework for connecting skill evaluation metrics with the results of a cybersecurity exercise.

2 Related Work

Although there is a large volume of published studies describing different aspects of cybersecurity exercises, there are not so many that would suggest a structured approach for their creation and implementation process. In this section we outline three examples of a structured approach. Illustration of those suggested processes is given in the Fig. 1 with each phase shown in their approximate position related to similar phases in other frameworks.

ENISA [4] outlines the process of cyber-exercises in four steps: (1) identifying objectives, scenario, type of exercise, key participants, size and scope; (2) planning; (3) conducting; (4) evaluating. While providing a good high-level overview of the topic, the guide by ENISA does not go into technical details of the implementation. It mentions the necessity to set SMART [5] objectives, but focuses more on questionnaires for the source of information. Skills required for successful implementation are mentioned as one of the possible objectives, but no further guidance is provided.

Patriciu et al. [6] suggest a seven step model: (1) objectives; (2) approach; (3) topology; (4) scenario; (5) rules; (6) metrics; (7) lessons learned. This model is linear and easy to follow. Suggested metrics include the time taken to achieve the technical goals, number of participants who successfully achieved technical goals and participants' satisfaction score from feedback forms. Compared to other models, it skips the execution phase of the exercise and only mentions steps for the preparation and follow-up activities.

Wilhelmson et al. [7] propose a six-step model: (1) idea generation; (2) preliminary investigation; (3) establishment - including defining mission statement, having a dialogue about the mission statement, planning the project and its organising process; (4) project initiation and implementation; (5) completion and reporting; (6) following-up and evaluation of results. In the evaluation phase, they focus on reflectively analysing the collected experiential feedback. An appointed evaluator is expected to present a report that includes a review of the exercise performance and lessons learned.

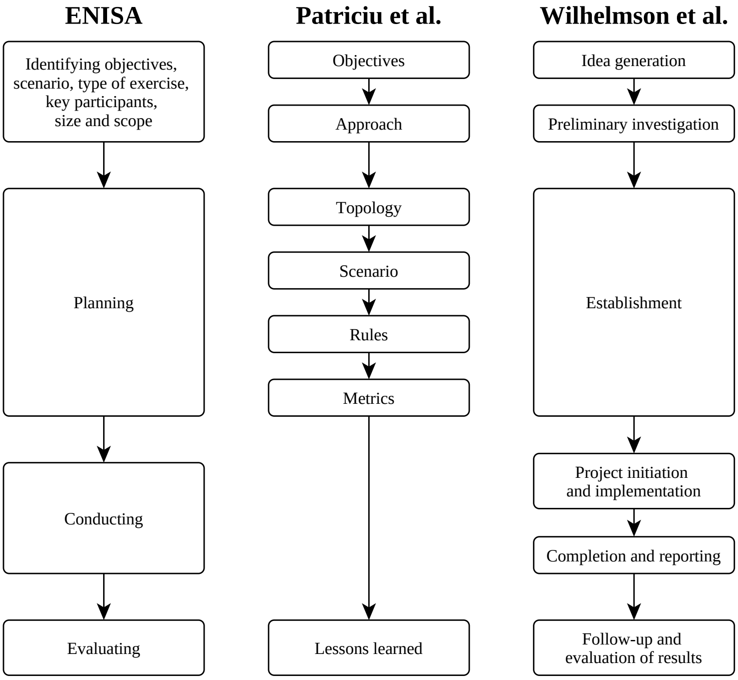

ENISA
Identifying objectives, scenario, type of exercise, key participants, size and scope
Planning
Conducting
Evaluating
Patriciu et al.
Objectives
Approach
Topology
Scenario
Rules
Metrics
Lessons learned
Wilhelmson et al.
Idea generation
Preliminary investigation
Establishment
Project initiation and implementation
Completion and reporting
Follow-up and evaluation of results

Fig. 1. Comparison of different suggested processes for conducting a cybersecurity exercise

Overall, the discussed models display similar underlying processes with a slightly different focus. However, these models pay little attention to measuring specific skills during the exercise. For example, Wilhelmson et al. mention that "the exercise participants can be involved in designing the exercise by sharing their ideas about what knowledge or skills in information and cyber security they would like to improve during the exercise" [7]. Nevertheless, they provide no specific further guidance for incorporating skill assessment into the exercise. Similarly, in other papers describing cybersecurity exercises, no evident attempt is made to evaluate the skills of participants in a way that would be understood outside of the context of the particular exercise. No guidance is provided for a structured evaluation of the skills of participants.

In addition to the papers that look at the organising part of cybersecurity exercises, there are others that take the first steps studying the related psychological aspects. For example, Katsantonis et al. [8] construct a concept map of the technological and pedagogical characteristics of live competitions. Bashir et al. [9] present the results of an extensive survey of cybersecurity competition participants providing a profile of the demographic, psychological, cultural, and vocations characteristics of competitions participants. Nevertheless, there seems to be a lack of holistic approach for fitting the psychological characteristics and technical process together.

The aim of the current paper is to introduce the developed holistic model for cybersecurity exercises that includes an integrated competency assessment. First, the new 10-step model is described. A separate section is devoted for a more detailed description of the process of measuring cybersecurity-related skills during the exercise (i.e. step 3 in the 10-step model). In addition, the described steps are illustrated using examples from the initial progress on developing CyberSpike [10] - a national cybersecurity exercise in Estonia.

3 Our Approach

Different cybersecurity exercises can have various forms, but most of them share some common characteristics. Looking at the general workflow of different approaches reveals that they seem quite well aligned as shown in Fig. 1. However, the simplified process descriptions might give a false impression of the workflow being linear. Based on our experience, the actual process of creating and implementing a cybersecurity exercise is much more iterative - similarly to the philosophy of agile software development. Therefore, instead of a linear process, we suggest to use a more flexible approach. The suggested 10-step model together with its iteration loops is shown in Fig. 2 and explained further in the following sections.

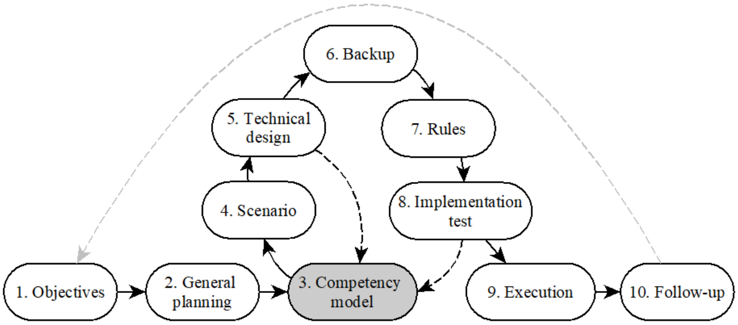


Fig. 2. Suggested 10-step process for designing and implementing a cybersecurity exercise

Fig. 2 shows how the design and implementation process of a cybersecurity exercise is usually not a straightforward process. After setting the high-level objectives (1) and the scope (2) for the exercise, it might take multiple iterations of work with competency model (3), scenario (4) and technical design (5) to reach a satisfactory result. Even after mitigating possible risks with backup plans (6) and setting the general rules

(7), the implementation test (8) might raise some bugs or concerns that require some part of the process to be remade. After a successful implementation test the actual exercise takes place (9) followed by the retrospective feedback session (10). Often, the exercise is not a one-time event but a part of a series. In that case, the follow-up session can provide valuable input for setting the objectives for the next time. Next sections provide firstly a short overview of each phase of the 10-step model. Afterwards, the competency model (phase 3) is described in more detail.

3.1 Objectives

Different approaches that were mentioned earlier for conducting cybersecurity exercises all agree that the process should start from setting general objectives and that makes perfect sense. The main objective of a cybersecurity exercise is usually to teach and/or to evaluate participants. These objectives often come together and are sometimes difficult to distinguish. After all, to confirm that learning has happened, an evaluation must be carried out before and after the learning process.

In addition to the main goal, there is often motivation to promote the field of cybersecurity. Various institutions are in constant hunt for talent that could be found during one of cybersecurity exercises. Many governments have the goal of educating more cybersecurity specialists and cybersecurity exercises can help to draw attention to this field.

In our case, the main goal of the CyberSpike competition was to find the top young cybersecurity talent. Also, it was considered to be important to promote the field of cybersecurity amongst the Estonian youth and to select the team for the international European Cyber Security Challenge [11].

3.2 General Planning

This phase of the exercise creation deals with budget, timeline and technology. It is necessary to understand the constraints and realistically evaluate how much it is feasible to do for achieving the general objectives. Also, it is good to plan the activities and resources early on. In this stage, the target group and type of exercise can be specified.

In our case, we decided to limit the target group to ages 14 to 30. The national cybersecurity competition CyberSpike was decided to be held in two phases. First, the preliminary round where all registered participants could remotely access the simulation in limited time. Secondly, the final round, where the participants with the best results from the preliminary round would physically gather in one location.

3.3 Competency Model

In this phase, the relevant competencies are identified and connected to the relevant tasks. This phase is covered in more details later in a separate section.

3.4 Scenario

In our suggested 10-step process, the scenario stands for both the technical scenario as well as for the storyline. The general flow of the exercise is addressed in this stage. To improve the engagement of participants, an established model for creating captivating game design narrative could be followed [12]. There are several models for creating game narrative brought out by Dickey [13] and Simpson [14], out of which the Vogler's hero's journey was used to shape the storyline of CyberSpike competition. Also, frameworks related to gamification can help to make participation of the exercise more enjoyable.

In our case, the Octalysis framework by Chou [15] helped to analyse the overall game experience and further improve the scenario. For example, we discovered that initially, the game narrative had rather low level of epic meaning and social influence mentioned in the Octalysis framework. Slight modifications of the narrative helped to improve those aspects without the need to redesign the technical challenges themselves.

3.5 Technical design

Constraints set by general limitations of time and expected skill level of the target audience help to shape the technical challenges for the exercise. Targeted skills and scenario help to further refine the possible tasks that are given to the participants. However, it might happen that some of the tasks would not be possible or feasible to add to the exercise. Also, some parts of the scenario might not be realistic. Therefore, it should be stressed that the creation of a cybersecurity simulation exercise should be iterative. As noted in Fig. 2, it could take multiple iterations for competency model, scenario and technical design to fit together well. By the end of the technical design, the exercise should have a specific network topology together with target machines and the supporting infrastructure.

3.6 Backup

For a successful project, different risks should be mitigated. If any part of the procedural or technical processes fails, then it is good to have a backup plan ready. The backup plan might include both technical and procedural backups for the case when some part of the system fails. Technical backups can include having additional hardware ready to support the exercise. Procedural backups can consist of plans how to postpone, downgrade or substitute some activities if needed.

3.7 Rules

Scoring rules take the input from the objective of the exercise. While the scoring mechanism should be already implemented, it is possible to modify it based on the focus areas. For example, service availability points might have a bigger weight than the ones of confidentiality breaches. Additionally, it is important to communicate relevant restrictions to the participants. For example, creating a denial of service attack might also influence the general infrastructure and might therefore be discouraged or strictly forbidden.

3.8 Implementation test

No software is perfect from the first run. Going through the implementation enables to notice any possible shortcomings. While it is expected that smaller tests are conducted already earlier in the process, it is important to carry out a full-scale system integration test to confirm that all parts of the designed system work as expected. Also, it could give valuable feedback about possible fine-tuning of the exercise.

In our case, the implementation test provided useful insights about the time required to complete the given tasks.

3.9 Execution

When all the testing has been successfully conducted, then the system is ready for the participants. The actual exercise can be run separately or as a part of a bigger event such as a security conference.

In our case, we decided to hold a cybersecurity-related seminar to be run in parallel with the final round of the competition. That enables people who are not competing in the exercise, but still interested in the topic, also take part of the event.

3.10 Follow-up

Collecting feedback is effective way to analyse the organising side of the exercise and it is vital in case there is any plan to continue this exercise or replay it another time. Feedback questionnaires and retrospective analysis meetings can be used to gather and discuss the lessons that were learned during the preparation and implementation of the exercise.

4 Competency Model

Competency is a fuzzy term as shown by Le Deist et al. [16] In the context of the current paper we consider competency a set of knowledge, skills and abilities and focus mainly on practical skills that are relevant in the field of cybersecurity and that can be measured by performance in virtual simulations. We first introduce a tool for identifying main skill categories - Cybersec-Tech window. Afterwards, we focus on the Stenmap framework that is designed to provide a more structured approach for connecting results of cybersecurity exercises with the skills that were demonstrated by participants.

4.1 Cybersec-Tech Window

Cybersec-Tech window (Fig. 3) is a simple tool for illustrating the division of different types of skills needed in the field of cybersecurity. It can be used for agreeing the array of skills that would be relevant to the exercise objectives and to find the most suitable format for the exercise.

The idea behind the Cybersec-Tech window is to divide skills into four categories based on two skill dimensions. The x-axis represents the scale from skills that are not cybersecurity specific to cybersecurity-specific skills. The y-axis represents the scale from non-technical to technical skills. It should be noted that the Cybersec-Tech window is rather arbitrary and is aimed to provide a classification bases for cybersecurity-related skills. We have found it useful in creating common understanding among people with diverse backgrounds, in identifying skills targeted in cybersecurity exercise.

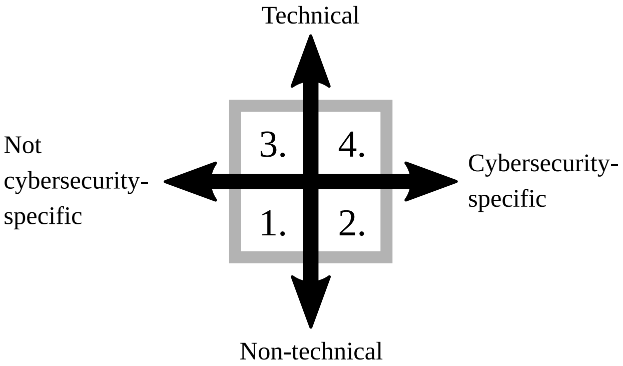


Fig. 3. Cybersec-Tech window

The axes of the Cybersec-Tech window divide it into following four quadrants:

1. Skills that are non-technical and not cybersecurity-specific. This is the place for universally transferable skills such as communication skills, leadership skills etc. While in our context, those skills are not considered technical, nor cybersecurity-specific, they are still highly valued. Tabletop exercises are often created to address this quadrant of Cybersec-Tech window. If the exercise is carried out in teams, then teamwork requires skills from this quadrant.
2. Second quadrant includes skills that are cybersecurity-specific, but non-technical. Security awareness trainings often target those skills. Not opening suspicious links in phishing emails or creating a secure password might not require very technical skills, but are nevertheless important to ensure a good security posture.
3. Third quadrant consists of skills that are technical, but not necessarily cybersecurity-related. Coding in some programming language is a good example of a skill belonging to this area.

4. Fourth quadrant consists of skills that are both technical and cybersecurity-specific. Implementing different encryption algorithms or performing a SQL injection attack are some samples of those type of skills.

It should be noted that it is not always easy to position a skill in this Cybersec-Tech window. For example, skills related to incident reporting could be general non-technical type or very specific and technical. Nevertheless, this Cybersec-Tech window can help to facilitate a discussion about which skills the cybersecurity exercise should target.

In our case, we decided to focus on quadrant 4 - i.e. technical skills that are cybersecurity-specific. It was acknowledged, that some skills from quadrants 1 and 2 might be required as well as several skills from quadrant 3. Fig. 4 illustrates the approximate area of skill types that were decided to be addressed in the exercise.

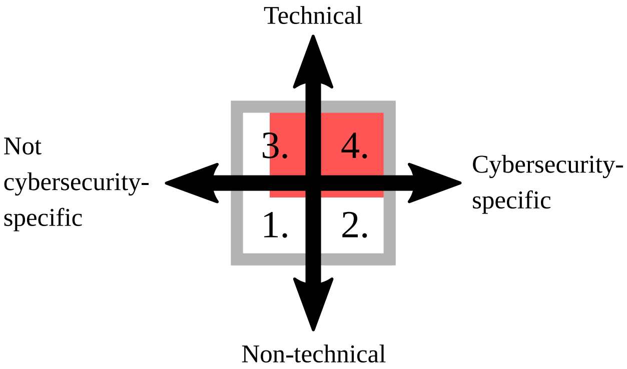


Fig. 4. Area of skill types that were decided to be addressed in the CyberSpike exercise

Cybersecurity exercise can have many formats. Some exercises are played in teams, others individually. Some simulations are played against other participants, others against the automated scenario. In the case of competing teams, it might not be possible to separate individual skills of the participants (unless the responsibilities are clearly and strictly divided between the team members). Still, the overall mapping of the skills of each team can be very valuable for identifying shortcomings that should be improved in the future.

In our case, we focused mainly on the quadrant 4 or Cybersec-Tech window - i.e. technical skills connected to cybersecurity. Some need for general system administration skills was identified. Based on the selected area of skills, we decided to hold an individual competition, where every participant is competing in an identical environment that is isolated from other participants. The Cybersec-Tech window helped to identify the area of skills to be focused on in the next phase using Stenmap framework.

4.2 Stenmap

In this section, the Stenmap[1] framework is described. Stenmap aims to map the relevant skills that can be measured by the serious game and then connect them to the results of the exercise. The main goal of Stenmap is to provide a general roadmap merging existing models in a structured way.

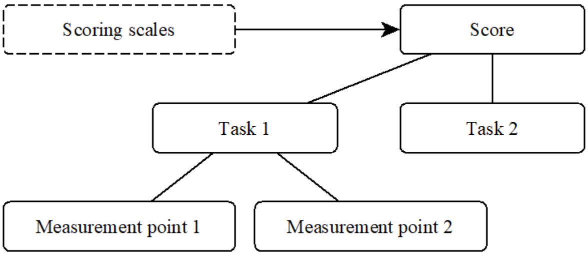


Fig. 5. Usual scoring system hierarchy

Fig. 5 illustrates a part of the usual scoring system of a cybersecurity exercise that consists of various tasks. For each successful completion of the task, a score is added to the final result. In more complex exercises, the scoring mechanisms might be more sophisticated and more important exercises can have a bigger weight and therefore a bigger impact to the final score. Evaluating the skills of the participant still requires expert interpretation based on the task completion metrics such as time and success ratio.

The goal for using Stenmap is to get a mapping of skills (possibly divided into areas or sets of skills) with according proficiency scales. It should be noted that using Stenmap is not excluding the possibility to still gather regular scores as well in order to find the winner of a competitive cybersecurity exercise. **Fig. 6** outlines the underlying hierarchy of Stenmap.

[1] Naming inspired by the infamous network mapping tool Nmap

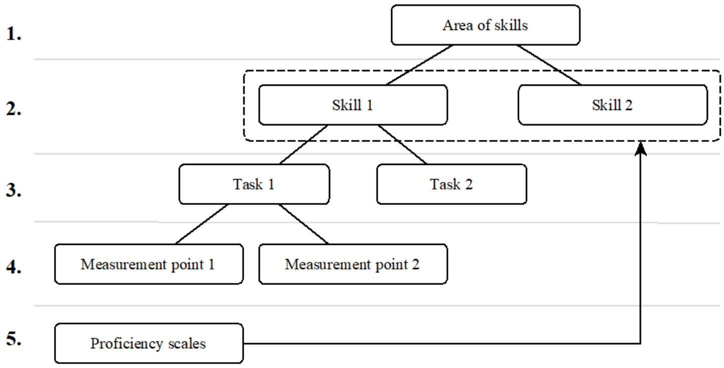


Fig. 6. Stenmap

The numbered levels 1. to 5. marked on the left side of **Fig. 6** represent the steps in the process of Stenmap as follows:

1. Identify relevant areas of skills (sets of skill);
2. Specify skills;
3. Specify tasks for each skill;
4. Identify suitable measurement point for each task;
5. Decide on proficiency rating scale and connect it to measurement points and/or tasks.

First, it is necessary to identify the area of skills targeted during the exercise. For establishing initial scope, the Cybersec-Tech window can be used as described earlier. Focus group interviews or discussions with specialists could also be beneficial for understanding the general skill set addressed by the exercise.

Steps 2 and 3 of Stenmap process deal with specifying skills and according tasks. For this purpose, a Job Performance Model (JPM) could be applied as described by Tobey [17]. According to Tobey [17], the JPM approach can be described in following steps:

1. Establish vignettes (or scenarios) that define situated expertise in job roles;
2. Detail the goals and objective metrics that determine successful performance;
3. Identify responsibilities by job role necessary to achieve the objectives;
4. Detail the tasks, methods, and tools along with how competence may differ by level of expertise or by the difficulty of achieving that level of expertise.

Alternatively, an existing framework could be used that includes relevant skills and task descriptions. Such frameworks are, for example, e-Skills Match described by Fernández-Sanz et al. [18] and NICE Cybersecurity Workforce Framework by NIST

[19]. In our case, we decided to use NIST NICE framework because it provided a ready-to-use list of skill areas, skills and tasks relevant to cybersecurity.

In the fourth step of Stenmap, the defined skills need to be connected to specific measurement point. As described by Mäses et al. [20] there can be multiple measurement points connected to one skill and vice versa.

Finally, a proficiency rating scale needs to be set and connected to the framework. The goal for proficiency scale is to give more information about a specific skill than just binary confirmation about the skill being present or not. It is possible to implement any scale, e.g. 0 to 4, 0 to 10 or even 0 to 127 as it has been done in some cases [21]. The levels of the scale could be explained by university grading system or follow some well-established taxonomy such as Bloom's taxonomy [22], N-loop learning levels [23] or SOLO taxonomy [24].

In our case, we decided to use Bloom's taxonomy because it is more focused on the task completion, contrary to SOLO taxonomy that requires more insight into the cognitive processes of the participant. In the future, we consider using SOLO taxonomy if we manage to collect more data about the way how participants complete their tasks.

To sum up, Stenmap enables to have a more in-depth analysis of the competition results - presenting not only the final ranking, but also an overview of the skills of the participants. Better measurement of existing cybersecurity-related skills enables also to create and validate systems for developing them.

5 Discussion and future work

Although various cybersecurity-related simulations have been conducted for quite some time, there is abundant room for further progress in identifying and analysing cybersecurity-related skills with the help of serious games. In the future work, the validity of the skill evaluation should be analysed by cross-validating the results of the competition - i.e. conducting additional evaluation of the participants. Ideally, the skill-based results of the cybersecurity exercises could be comparable between different exercises. Considerably more work will need to be done to achieve such comparability. Implementing Stenmap to move from simple score-based results to a more meaningful evaluation of skills is only the first step on that path.

Another goal for the further perspective would be the opportunity to draw some general conclusions based on the structured data in Stenmap. For example, it might be possible to infer from a person who can properly configure a particular firewall, that he/she is likely to be able to set up a web server. This kind of connections between different competencies require more experimental data and further investigation.

6 Conclusion

In this work we have addressed the problem of evaluating cybersecurity-related skills with the help of computer simulations. Building upon previously suggested process models for creating and implementing a cybersecurity exercise, a new 10-step approach was presented. As the actual process of creating and implementing a cybersecurity exercise was found to be much more iterative than indicated by previous research, the new 10-step model includes iteration loops for describing a more agile approach. Very little was found in the literature on systematic measuring of specific skills during the cybersecurity exercises. Although it is possible to interpret the results of a simulation to evaluate the skills of the participants, it is not being done in a standardised and scalable way. To advance this situation, the current paper introduces a conceptual theoretical framework assisting future work on skill evaluation and improvement.

Special attention was given to the phases of the 10-step model that were connected to skill evaluation. Namely, the Cybersec-Tech window was introduced as a tool for discussing the high-level classification of cybersecurity-related skills and reaching a common understanding among exercise organisers coming from diverse backgrounds. Additionally, Stenmap framework was described for providing a structured approach for skill evaluation in cybersecurity simulations. We showed how the focus area identified using the Cybersec-Tech window can be further specified and connected to relevant tasks. Also, we mentioned some suitable taxonomies that would provide a skill profile that would be comparable to results of other similar exercises. Overall, the mentioned parts together form the necessary basis for creating cybersecurity simulations, that provide insights to the skills of participants in a way that is comparable not only between participants of one event, but also between different exercises.

Acknowledgements

This research was supported by the European Regional Development Fund of European Union and Estonian Ministry of Defence.

References

1. Conti, G., Babbitt, T., Nelson, J.: Hacking competitions and their untapped potential for security education. IEEE Security and Privacy. 9, 56–59 (2011).
2. Knowles, W., Such, J.M., Gouglidis, A., Misra, G., Rashid, A.: All That Glitters Is Not Gold: On the Effectiveness of Cyber Security Qualifications. 1–10.
3. Furnell, S., Fischer, P., Finch, A.: Can’t get the staff? The growing need for cyber-security skills. Computer Fraud and Security. 2017, 5–10 (2017).
4. ENISA: Good Practice Guide on National Exercises. 80 (2009).
5. Bjerke, M.B., Renger, R.: Being smart about writing SMART objectives. Evaluation and Program Planning. 61, 125–127 (2017).
6. Patriciu, V., Furtuna, A.C.: Guide for Designing Cyber Security Exercises 2 . The Need for a Uniform Structure. Proceedings of the 8th WSEAS International Conference on E-Activities and Information Security and Privacy. 172–177 (2009).
7. Wilhelmson, N., Svensson, T.: Handbook for planning, running and evaluating information technology and cyber security. (2013).
8. Katsantonis, M., Fouliras, P., Mavridis, I.: Conceptual Analysis of Cyber Security

Education based on Live Competitions. 771–779 (2017).

9. Bashir, M., Lambert, A., Wee, J.M.C., Guo, B.: An Examination of the Vocational and Psychological Characteristics of Cybersecurity Competition Participants. USENIX Summit on Gaming, Games, and Gamification in Security Education. 1–8 (2015).
10. CyberSpike, http://cyberspike.ee.
11. ENISA: European Cyber Security Challenge, https://www.europeancybersecuritychallenge.eu.
12. Chothia, T., Holdcroft, S., Radu, A., Thomas, R.J.: Jail, Hero or Drug Lord? Turning a Cyber Security Course Into an 11 Week Choose Your Own Adventure Story. 2017 USENIX Workshop on Advances in Security Education (ASE 17). (2017).
13. Dickey, M.D.: Game design narrative for learning: Appropriating adventure game design narrative devices and techniques for the design of interactive learning environments. Educational Technology Research and Development. 54, 245–263 (2006).
14. Simpson, J., Coombes, P.: Adult learning as a hero's journey: Researching mythic structure as a model for transformational change. Queensland Journal of Educational Research. 17, 164–177 (2001).
15. Chou, Y.: Actionable Gamification - Beyond Points, Badges, and Leaderboards. Octalysis Media (2015).
16. Le Deist, F.D., Winterton, J.: What is competence? Human Resource Development International. 8, 27–46 (2005).
17. Tobey, D.H.: A vignette-based method for improving cybersecurity talent management through cyber defense competition design. In Proceedings of the 2015 ACM SIGMIS Conference on Computers and People Research. New York, NY: ACM. Proceedings of the 2015 ACM SIGMIS Conference on Computers and People Research - SIGMIS-CPR '15. 31–39 (2015).
18. Fernández-Sanz, L., Gómez-Pérez, J., Castillo-Martínez, A.: e-Skills Match: A framework for mapping and integrating the main skills, knowledge and competence standards and models for ICT occupations. Computer Standards & Interfaces. 51, 30–42 (2017).
19. Newhouse, W., Keith, S., Scribner, B., Witte, G.: National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework. NIST Special Publication 800-181. (2017).
20. Mäses, S., Hallaq, B., Maennel, O.: Obtaining Better Metrics for Complex Serious Games Within Virtualised Simulation Environments. ECGBL 2017 11th European Conference on Game-Based Learning. 428–434 (2017).
21. Umbleja, K.: Competence Based Learning Framework , Implementation , Analysis and Management of Learning Process, (2017).
22. Starr, C.W., Manaris, B., Stalvey, R.H., Starr, C.W., Manaris, B., Stalvey, R.H.: Bloom's taxonomy revisited. ACM SIGCSE Bulletin. 40, 261 (2008).
23. Simonin, B.L.: N-loop learning: part I – of hedgehog, fox, dodo bird and sphinx. The Learning Organization. 24, 169–179 (2017).
24. Brabrand, C., Dahl, B.: Using the SOLO taxonomy to analyze competence progression of university science curricula. Higher Education. 58, 531–549 (2009).

Appendix 3

III

M. Ernits, K. Maennel, S. Mäses, T. Lepik, and O. Maennel. From simple scoring towards a meaningful interpretation of learning in cybersecurity exercises. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security: ICCWS 2020*. Academic Conferences and publishing limited, 2020

From Simple Scoring Towards a Meaningful Interpretation of Learning in Cybersecurity Exercises

Margus Ernits[1], Kaie Maennel[2], Sten Mäses[2], Toomas Lepik[2] and Olaf Maennel[2]
[1]RangeForce.com, Tallinn, Estonia [2]School of Information Technologies, Tallinn University of Technology, Tallinn, Estonia

margus.ernits@rangeforce.com
kaie.maennel@taltech.ee
sten.mases@taltech.ee
toomas.lepik@taltech.ee
olaf.maennel@taltech.ee

Abstract: To overcome the current skills shortfall in cybersecurity, a broad range of IT professionals and users should be educated in the fundamentals of protecting computer systems and the data they contain. This requires novel and scalable teaching methods. The main contribution of this paper is to introduce an approach of how to create cybersecurity exercises that can measure relevant competencies. We demonstrate how technical event logging can be linked to learning outcomes and skills measurement by defining intermediate abstraction layers. These take raw forensic data from the game-system and network, and gradually group them into events and abstract measurements until they can be mapped to learning theories. The suggested approach enables deeper insights into learning. This approach has been applied for developing labs using our cloud-based open-source tool. The labs have been used by more than 2 000 learners in over 15 000 sessions. A thorough hands-on skills assessment was conducted before and after a set of exercises for 27 participants. Results show that the suggested method can be used for creating and improving cybersecurity exercises.

Keywords: cybersecurity education, exercises, skill assessment, learning

1. Introduction

Cybersecurity exercises are becoming increasingly popular for educating and evaluating security specialists (Ogee et al, 2015). This comes as no surprise when our society is fundamentally dependent on IT systems and vulnerabilities are subject to exploitation by threat actors. To overcome today's cybersecurity problems, a very broad range of IT professionals should be educated and everyone should understand the fundamentals of cybersecurity.

Therefore, novel and scalable teaching methods need to be developed. Realistic attacks and complex simulated systems in virtualized environments can provide engaging and practical hands-on learning experiences using fully automated training that utilizes adaptive learning methods. The goal is to have a training environment that would detect the skill-level of the learner and automatically select the most appropriate learning tasks for the user. Thus, it is important to have a measurement methodology that is able to accurately capture the capabilities of the user. However, current research suggests (e.g., Fulton et al, 2012) a lack of defined educational outcomes. This might be due to the overall difficulty of designing and implementing a complex defence oriented gamified cybersecurity exercise. Specifically, constant adjustments to the scoring system and storyline are usually very time consuming and may divert the attention from in-depth analysis of the final score.

Our aim is to apply the existing instructional design methods for connecting raw data points to high level competencies using an evidence-correlation model. Specifically, we suggest a method for the design and implementation of an exercise that would give a structured and automatic feedback of the participants' skills and competencies. This is implemented on i-tee, an open-source software platform (Ernits and Kikas, 2016) developed from experience gained in several large-scale exercises including Locked Shields and Cyber Security Challenge UK. Nevertheless, the suggested model is itself platform independent and can be implemented using other environments such as OpenStack.

Over 2 000 learners have used the system in more than 15 000 various lab sessions on a wide range of topics. Of those, the skills of a pilot group of 27 IT developers were measured before and after a set of training exercises.

The set consisted of 13 different labs: Command Injection, Cookie Security: Secure, Cookie Security: HttpOnly, Cross-Site Request Forgery (CSRF), Defence against CSRF, Insecure Direct Object Reference, Intro Lab, Path Traversal, SQL Injection, Unrestricted File Upload, Reflected Cross-Site Scripting (XSS), Stored XSS and Phishing based on Stored XSS.

2. Related work

Performance measurement is crucial for mastering a skill as argued in *Understanding by Design* by Wiggins (2005). There is also a growing body of work on this topic in the field of cybersecurity exercises. This section focuses on existing conceptual approaches and evaluation models in cybersecurity training, with a focus on training through exercises.

2.1 Design approaches to gamified cybersecurity training

Katsatonis et al (2017) provide a concept map of cybersecurity game-based approaches' key elements that include also learning objectives and assessment. Learning objectives should be based on performance, proficiency and be connected to game-play. Vykopal et al (2016) suggest decomposing the training activity into individual levels that learners have to accomplish for satisfying specific learning objectives. The data collected include metrics such as the start and end of each game and level within it. More detailed data include submission of incorrect flags and their content, hints used, skipping a level, displaying a level's solution and game ID (Vykopal et al, 2016). Vykopal et al (2017) state that setting learning objectives based on learners' skills before the actual exercise is a challenging undertaking.

Clark (2015) proposes a 4-level model, where in order to create a broader understanding of the security elements the impact on the target-host, server, firewall and intrusion detection systems is highlighted at each level. Nicholson et al (2016) suggest that learning should be individualised to fit the higher skill or competency level the subject is aiming for. These should consider the learners' profiles, content brokering, experience tracking and competency network. With these high-level conceptual approaches our proposed model—experience and competency tracking allows the learner profiles to be updated based on the progress in competencies and also need to be easily adjusted.

2.2 Evaluation models used in cybersecurity exercises

There are multiple frameworks looking at classification of cybersecurity-related competencies. The National Initiative for Cybersecurity Education Cybersecurity Workforce Framework (NICE Framework), published by the National Institute of Standards and Technology (Newhouse et al, 2017) lists knowledge, skills and abilities required to perform tasks in specific work roles. Rashid et al (2018) introduce the Cyber Security Body of Knowledge project aiming to codify the foundational and generally recognized knowledge on cybersecurity. Nevertheless, those high level frameworks provide only generic ideas instead of specific tasks for measuring skills. For example, task T0349 in NICE framework is described as "collect metrics and trending data" (Newhouse et al, 2017) leaving space for various interpretations about the task specifics.

Abbott et al (2015a) provide a quantitative evaluation of techniques for student performance assessment. This uses an automated mechanism for parsing log entries into blocks of time during which participants are focused on specific high-level objectives. Abbott et al (2015b) also describe an exercise instrumentation that enables automated performance assessment by capturing students' computer-based transactions in a log. This is time-synced with the game-server to deliver challenges and registers student responses. Labushange and Grobler (2017) describe assessing technical skill level based on indexed similarity of the commands used to achieve the specified objectives from which the level of participant's practical knowledge could be inferred. The paper classifies learner's actions that can automatically deducted using the clustering of commands but does not go into details. The similar underlying idea is implemented in our model at raw data to Game Event Logs (GEL) transition level.

Many articles address different issues in cybersecurity exercises including skills assessment and evaluation attempts. Scoring is often seen as a tool to provide evaluation and feedback to exercise participants. However, scoring systems are not necessarily connected to learning objectives. What is missing is a practical and scalable model that would provide evidence that high level competencies can be achieved through analysing the granular data level of the exercises' raw data.

3. Connecting competences to raw data

In cybersecurity exercises, different events happen rapidly in a semi-controlled environment. However, learning experiences are not linear or predictable. For example, in case the participants have a task to defend a vulnerable web application, they may have different correct ways to deal with the attacks. These include block attacks by implementing intrusion prevention systems by using web application firewalls or by fixing the vulnerabilities of web application. There are also different incorrect or insufficient ways to react such as removing the attacker's injected code/content from the system, taking vulnerable applications offline or by breaking web application's functionality. These make measuring specific competencies challenging.

The primary focus when designing an exercise is to start with the conceptual design. First, targeted competencies that the exercise should teach or assess are defined. These competencies are decompiled to different skills with measurable learning objectives. Based on those learning objectives, specific tasks can be determined that could be measured by evidence—i.e., events happening in the system.

To improve the flexibility and efficiency of the system, Game Event Logs (GEL) are designed to capture all the important events. When GEL are designed appropriately, the amount of the exercise data needed to analyse decreases significantly. Massive amounts of raw data can be deleted after the exercise while maintaining the ability to dynamically change the rules for interpreting the events. The raw data is used to generate GEL that are interpreted to evaluate whether a specific task is completed. Commonly, the completion of tasks is used as an input for score calculation, but we suggest doing more than that. The completion of tasks indicates the proficiency level of different skills. Skills in turn are gathered into meaningful sets to form competencies.

Fig. 1 illustrates the suggested design model for exercises enabling the evaluation of participant competencies. The model consists of 5 layers with the bottom layer representing the raw data and the highest layer being specific competencies that are targeted by the exercise. It is important to note that while the technical data flows from the 1st to the 5th layer, the logical design flow should start from the top layer. The layers themselves are connected to each other but at the same time independent, allowing the use of different formats or processing systems. In the following sections, we explain each layer in detail and use Cross-Site Scripting (XSS) related competency for illustration.

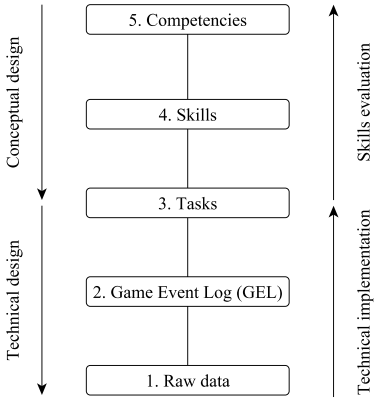


Figure 1: Suggested structure for exercises

3.1 Layer 5 - Competencies

The creation of every exercise should start with question "Why?". In a cybersecurity context, it is usually the need to protect the systems or build more secure systems and to achieve this goal, organisations need people with specific competencies. The term competency is somewhat ambiguous as shown by Le Deist and Winterton (2005). In the context of this paper, we define competency as a set of knowledge, skills and abilities. These focus mainly on practical skills that are relevant in the field of cybersecurity and that can be measured by performance in virtual hands-on exercises.

The competency can be taken as a general theme for a set of skills that it encompasses. For example, defending against Cross-Site Scripting (XSS) was one of the competencies used in our labs.

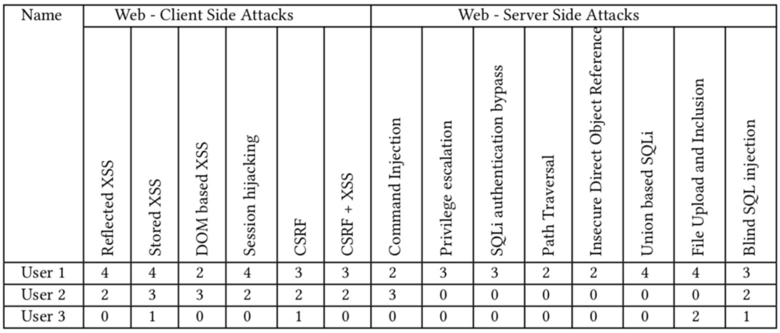
3.2 Layer 4 - Skills

Using competency as the general theme, we can look at skills as sub-topics of an area. As different jobs require groups of skills, it is possible to define the set of skills for a particular position and then evaluate an individual accordingly. However, it should be noted that dividing job qualifications (professionalism) into different skills (and sub-skills) is a subject of debate (Rigby and Sanchis 2006). In cybersecurity, the relevant skills and sub-skills can be further analysed based on types of attack vectors. In the XSS lab for example, the types of XSS attacks can be defined as reflected XSS, stored XSS, and DOM based XSS (Gupta and Gupta (2017). In our example, the exercise focuses on developing skills to defend against the reflected XSS.

The layer of skills defines both high- and low-level skills. High level skills are more meaningful and are used in daily conversations. Installing WordPress to a web server from scratch is an example of a higher-level skill. Lower level skills are usually not meaningful or useful by themselves. Being able to remotely log in to a server using SSH is an example of a sub-skill. It is not that useful by itself, but it is a needed step in a larger process. Note that skills and sub-skills can be used to specify different proficiency levels. Some sub-skills can be considered necessary for some higher-level skill and different learning objectives can indicate different skill proficiency levels.

Specifying a skill in a measurable way is not an easy task. For example, a learning objective can be 'Learner can fix a web application with a XSS type vulnerability'. An undesired fix, which would fulfil the task could be to take the site offline. This means that appropriate learning objective should be: 'Learner can fix a web application with a XSS type vulnerability without disturbing service or breaking functionality'.

Skills are closely related to learning outcomes. Learning objectives define the expected goal of exercise in terms of demonstrable skills or knowledge acquired by a participant as a result of exercise (Malan, 2000). The skills and knowledge can be analysed using different models such as Bloom's, SOLO, etc. Figure 2 illustrates how in our cybersecurity exercise training model we follow and map the task with a range of cognitive learning and skills layers. These are Not graded (0), Remembers (1), Understands risks/attacks (2), Applies attacks (3), Applies defences (4) and Masters defences (5).



Name	Web - Client Side Attacks						Web - Server Side Attacks							
	Reflected XSS	Stored XSS	DOM based XSS	Session hijacking	CSRF	CSRF + XSS	Command Injection	Privilege escalation	SQLi authentication bypass	Path Traversal	Insecure Direct Object Reference	Union based SQLi	File Upload and Inclusion	Blind SQL injection
User 1	4	4	2	4	3	3	2	3	3	2	2	4	4	3
User 2	2	3	3	2	2	2	3	0	0	0	0	0	0	2
User 3	0	1	0	0	1	0	0	0	0	0	0	0	2	1

Figure 2: Skills report

The cognitive learning layers based on the revised Bloom's taxonomy (Krathwohl and Andersen 2001) are remembering, understanding, applying, analysing, evaluating, and creating. The adjustment is made in order to incorporate the attack and defence aspects. In order to defend, the learner needs to understand and apply the attack technique themselves before being able to creatively avoid such vulnerabilities in the system. Difference in applying and mastering the defence is transferability such as assessing if the learner is able to defend against this type of attacks in different operating systems or tools. Mastering the defence (level 5) in our model usually means that the same skill is measured and mapped over different labs. This ensures that the learner is able to transfer and apply the skill using different technologies.

In our lab example, the learning objectives are as follows:

- Learner understands impact of the reflected XSS.
- Learner mitigates the attack (e.g., applies HttpOnly flag).
- Learner fixes XSS vulnerabilities (using PHP in this lab).
- Learner performs reflected XSS.
- Learner recognises reflected XSS.

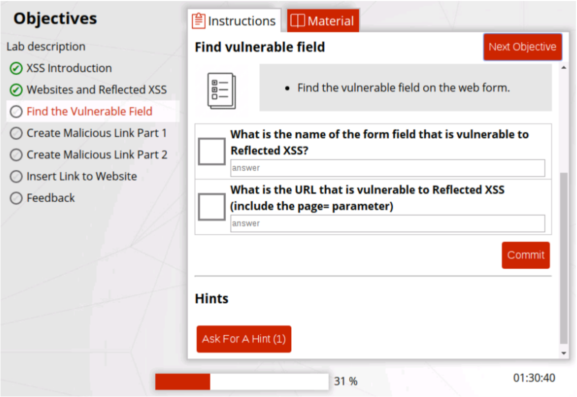
3.3 Layer 3 - Tasks

Tasks represent assignments that can be clearly defined and measured. Different tasks should represent different proficiency levels for the connected skills. For example, performing a blind SQL injection can be considered more advanced than a simple SQL injection.

In the reflected XSS lab, the tasks mapped to the cognitive learning layers are the following:

- Learner finds reflected XSS from unfamiliar target (Linux, php, web application).
- Learner uses reflected XSS found to retrieve session cookie of emulated computer user.
- Learner uses session cookie to login.

Tasks can be arranged in groups or in hierarchies. Fig. 4 shows an example from our XSS lab displaying a task of finding a vulnerable field in a web form. This task has two sub-tasks that are evaluated based on user input.



Find vulnerable field

- Find the vulnerable field on the web form.

□ What is the name of the form field that is vulnerable to Reflected XSS?
answer

□ What is the URL that is vulnerable to Reflected XSS (include the page= parameter)
answer

Figure 4: Example of tasks in the XSS lab

Tasks can be measured differently—by using direct input from the learner or with the help of some automation (Mäses et al, 2017). In case of automated task evaluation, it is useful to have a dedicated abstraction level for GEL.

3.4 Layer 2 - Game Event Log (GEL)

As seen in Fig. 1, the conceptual design of an exercise finishes with defining tasks. The exercise organisers do not need to go deeper into technical design. Therefore, starting from Layer 2, the focus moves from design to measurement. The question is how to quantify a particular task. This could be done by observing system logs or using specific scoring scripts such as a script that pings a web server and registers responses.

GEL enables flexible and meaningful rules for task evaluation. GEL include may references to the system log. For the XSS lab, the following events have been defined:

- The lab is personalised correctly (learner has successfully initialised the lab).
- Target website in the lab works correctly (based on user emulation checks).
- Learner has found the vulnerable form field (submitted correct field name without submitting all/lots of them).

In the following example there are two events described in GEL. The first one is stating that the simulated user (searching for the link with XSS injection) was active. The second example comes from a lab dealing with drones, where drone number 10 is functional, has operational backdoor (backdoor vulnerability not fixed), has XSS vulnerability (status false, because vulnerability is not fixed).

```
Apr 30 09:24:58 GEL index.js[926]: debug: Stored XSS: Simulating manager user

Apr 30 09:55:59 GEL index.js[1763]: info: [Drone 10] Status: functional=true,
backdoor=false, xss=false
```

GEL allows new rules to be added to higher levels of the model. These define new competencies, skills and tasks based on already existing game events, enabling rules to be redefined and fine-tuned more dynamically.

We use event correlation to evaluate objectives based on event logs. The attack evidence, such as successful or unsuccessful attacks, is collected and correlated with evidence that emulated computer users are able to use a full functionality of the simulated system.

3.5 Layer 1 - System state, raw data

The system state can be found from system logs or by a specific script checking such as whether the ping packet to the server gets a reply. Based on the system states, the event logs are created.

4. System architecture

The system architecture gives a more holistic view for understanding the suggested approach. However, the proposed method that links log events and raw data to learning outcomes does not depend on the underlying software stack. **The main goal is to measure the cyber skills in a scalable way.** Such measurement is a prerequisite for individually adaptive learning and enables to measure training effectiveness.

In our architecture, to access the hands-on exercise platform, a participant needs HTML5 capable web browser without any additional plug-ins or Virtual Private network(VPN). The system uses a Virtual Machine (VM) Host platform based on the open source tool i-tee (Ernits and Kikas, 2016). When a lab starts, all VMs, networks, and grading systems are provisioned and personalised for the participant. Also, the automated skill evaluation process is initialised. Each lab may include different VMs (Linux, Windows, BSD, etc.) and different software defined networks. Some VMs are accessible for participants (blue systems), whereas other VMs are dedicated for attack traffic generation (red systems) or for end-user simulation and for network traffic generation.

Fig. 4 illustrates the general architecture. The system provides network isolation between participants and lab networks. Lab personalisation process creates flags, vulnerabilities, grading for each lab attempt and random IP addresses for attacks and grading. Those IPs are based on real logs from our servers (fail2ban, sshguard, blacklists). The system provides interactive assistance and guidance for participants using hints, leaked hacker's chat live stream, and media injects using Virtual Teaching Assistant for the learner. Gamification elements such as leader-boards, scoreboards and hackers chatrooms are provided.

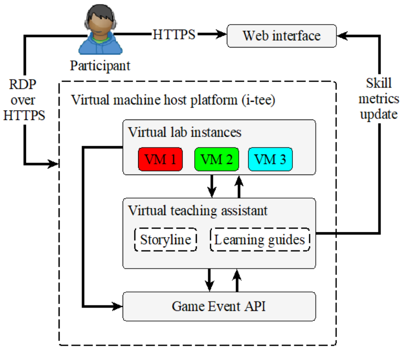


Figure 4: System Architecture

This architecture allows the creation of new labs and challenges by reusing existing modules such as attacking and assessment scripts and vulnerable targets. The system is designed to enable a lab to start without extra management effort. All game services, routers, networks, scoring bots are allocated on demand.

Our system scales as follows:

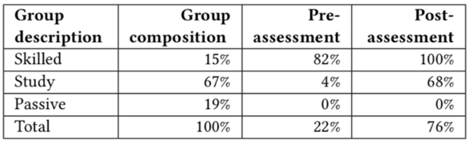
- First, we can easily teach small groups of students (1-30 lab sessions per server). When using cloud resources, we can run defence-oriented exercises with 300 participants without building a new expensive cyber range.
- Second, automatic assessment system demands no human red team (attacking team) or assessors.
- Third, the system is remotely accessible using web browser and can be used from home or in a classroom at any time.

5. Initial results

More than 2 000 learners have used the system in more than 15 000 lab sessions. We have used our platform for on-site cybersecurity competitions in 10 countries, in companies and academic training programs. A more thorough hands-on assessment was conducted with 27 participants as follows:

- Participants completed hands-on pre-assessment labs of approximately a 4 hours long assessment covering wide range of cybersecurity skills.
- Participants were assigned 13 different labs including XSS, command injection, cookie security.
- Participants completed hands-on post-assessment labs.

The results in Fig. 5 show average completion rate of pre- and post-assessment labs. The results are presented in three sub-groups. "Skilled" (high pre-assessment score, successful lab completions and post-assessment), "study" (low pre-assessment score, successful lab completions and post-assessment) and "passive" (low pre-assessment score and unsuccessful in lab completion or post-assessment). These are subjective sub-categories to analyse participants based on pre-assessment results and completion of the labs. It can be seen, for example, that for the "study" group, the improvement in skills was significant (from 4% to 68%).



Group description	Group composition	Pre-assessment	Post-assessment
Skilled	15%	82%	100%
Study	67%	4%	68%
Passive	19%	0%	0%
Total	100%	22%	76%

Figure 5: Skills improvement

After the exercise, additional free-form feedback was asked from the participants and from their supervisors to analyse whether participants had acquired relevant competencies. As anecdotal evidence, the organisation identified a participant who lacked cybersecurity-related experience and skills before training program but was able to identify 5 security vulnerabilities from their internal system after training. Such unstructured feedback has given subjective evidence of skill improvement, quantitative surveys could be used to gather feedback systematically.

6. Discussion

There are multiple initiatives towards more competency-driven computer science education that aim to focus more on competencies instead of primarily covering topics (Sabin et al, 2018). There are also discussions on topics such as integrating hands-on cybersecurity exercises into the curriculum (Weiss et al, 2019), and what core cybersecurity skills students should learn (Jones et al, 2018). The big question is: how to measure those skills?

Presenting learning outcomes without a clear mapping to relevant measurement points makes it difficult to evaluate individual skills. For example, a student who completes a course with a passing rate of 82% might be equally above average in all the relevant skills or have mastery of most skills and complete lack of others. Having a clear line of thought between different abstraction levels helps to provide evidence for the achieved competency.

From the learner's perspective, the main added value of our competency-driven approach is automated (therefore timely) individual feedback regarding the learner's skills. The individual skills report (such as illustrated by Fig. 2) enables the learners to find their current strengths and weaknesses and modify their future learning activities accordingly. For the evaluator, the skills report enables to track learners' progress focusing on their actual skills. At the same time, automated and scalable approach helps to reduce the evaluator's workload.

Following the competency-driven design can seem like a challenging task at first. However, the effort pays off because it helps to keep a clear overview when modifying or scaling up the system. Also, such competency-driven structure helps to remove the tendency to design too simple and easy-to-create tasks and potentially not covering intended core skills. Exercise design is an iterative process (Mäses et al, 2018) and technical design can influence task selection (e.g., technical implementation might be too complex or just unfeasible). Additionally, there are ethical considerations, which might force to abandon measurement of some initially planned skills. Exercises are often aimed to be as realistic as possible (Fox et al, 2018), but in the design phase a balance should be struck between what is realistic and what is measurable. In measuring a skill, the realism of a task is not the ultimate goal and it can be tailored to make the task quantifiable.

A more universal competency evaluation forces the educators to focus on how to measure higher level skills in a scalable way.

7. Future Work

Our current work focuses on defining the rules that can be universally applied in different exercises irrespective of the platform. This enables the organisers to dive into the learning data more easily. This data could be used for modelling predictive behaviours, such as considering the use of learning hints in skills' evaluation (Chow et al, 2017) and calculating confidence correlation to the skills. Our method can be used for connecting raw data to widely used competency frameworks such as NICE Framework. Potentially, it could be also used for comparing the learners' profiles to current job market requirements.

6. Conclusion

The opportunity to benchmark the competencies in a comparable way provides more insights that simply offering scores from cybersecurity exercises. Our structured approach helps to obtain more meaningful learning data from the logged events instead of counting points. Our main contribution is demonstrating how to create cybersecurity exercises that measure relevant competencies. We have applied this method for developing labs and assessing the skills of 2000 learners (including 27 having a more thorough assessment). Initial validation results show that connecting cybersecurity exercise raw data to the skills and competencies is a promising way forward.

Our approach supports a paradigm shift towards the cybersecurity exercises that by design allow the systematic and evidence-based competencies and skills measurement. The open-source platform, described design and evaluation process with ultimate aim to measure competencies, form together a step to achieve this change.

References

Abbott, R.G., McClain, J., Anderson, B., Nauer, K., Silva, A. and Forsythe, C., 2015. Log analysis of cyber security training exercises. Procedia Manufacturing, 3, pp.5088-5094.

Abbott, R.G., McClain, J.T., Anderson, B.R., Nauer, K.S., Silva, A.R. and Forsythe, J.C., 2015. Automated Performance *Assessment in Cyber Training Exercises*. Sandia National Lab, Albuquerque, NM.

Chow, S., Yacef, K., Koprinska, I. and Curran, J., 2017, July. Automated data-driven hints for computer programming students. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 5-10). ACM.

Clark, D.J., 2015, November. An onion approach to cyber warfare training. In *2015 Military Communications and Information Systems Conference* (pp. 1-4). IEEE.

Ernits, M. and Kikkas, K., 2016, July. A live virtual simulator for teaching cybersecurity to information technology students. In *International Conference on Learning and Collaboration Technologies* (pp. 474-486). Springer, Cham.

Fox, D.B., McCollum, C.D., Arnoth, E.I. and Mak, D.J., 2018. Cyber Wargaming: Framework for Enhancing Cyber Wargaming with Realistic Business Context.

Fulton, S., Schweitzer, D. and Dressler, J., 2012, October. What are we teaching in cyber competitions?. In *2012 Frontiers in Education Conference Proceedings* (pp. 1-5). IEEE.

Gupta, S. and Gupta, B.B., 2017. Cross-Site Scripting (XSS) attacks and defense mechanisms: classification and state-of-the-art. *International Journal of System Assurance Engineering and Management*, *8*(1), pp.512-530.

Jones, K.S., Namin, A.S. and Armstrong, M.E., 2018. The core cyber-defense knowledge, skills, and abilities that cybersecurity students should learn in school: Results from interviews with cybersecurity professionals. *ACM Transactions on Computing Education*, *18*(3), p.11.

Katsantonis, M.N., Fouliras, P. and Mavridis, I., 2017, September. Conceptualization of Game Based Approaches for Learning and Training on Cyber Security. In *Proceedings of the 21st Pan-Hellenic Conference on Informatics* (p. 36). ACM.

Krathwohl, D.R. and Anderson, L.W., 2009. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Labuschagne, W.A. and Grobler, M., 2017, June. Developing a capability to classify technical skill levels within a Cyber Range. In *ECCWS 2017 16th European Conference on Cyber Warfare and Security* (p. 224). Academic Conferences International Limited.

Le Deist, F.D. and Winterton, J., 2005. What is competence?. *Human resource development international*, *8*(1), pp.27-46.

Malan, S.P.T., 2000. The 'new paradigm' of outcomes-based education in perspective. *Journal of Consumer Sciences*, *28*(1).

Mäses, S., Hallaq, B. and Maennel, O., 2017, October. Obtaining better metrics for complex serious games within virtualised simulation environments. In *European Conference on Games Based Learning* (pp. 428-434). Academic Conferences International Limited.

Mäses, S., Randmann, L., Maennel, O. and Lorenz, B., 2018, July. Stenmap: framework for evaluating cybersecurity-related skills based on computer simulations. In *International Conference on Learning and Collaboration Technologies* (pp. 492-504). Springer, Cham.

Newhouse, W.D., Keith, S., Scribner, B., Witte, G., 2017. *Nice cybersecurity workforce framework: National initiative for cybersecurity education* (No. Special Publication (NIST SP)-800-181).

Nicholson, D., Massey, L., O'Grady, R. and Ortiz, E., 2016. Tailored Cybersecurity training in LVC environments. In *MODSIM World Conference, Virginia Beach, VA*.

Ogee, A., Gavrila. R., Trimintzios, P., Stavropoulos, V. and Zacharis, A. [n. d.]. The 2015 Report on National and International Cyber Security Exercises.

Rashid, A., Danezis, G., Chivers, H., Lupu, E., Martin, A., Lewis, M. and Peersman, C., 2018. Scoping the cyber security body of knowledge. *IEEE Security & Privacy*, *16*(3), pp.96-102.

Rigby, M. and Sanchis, E., 2006. The concept of skill and its social construction. *European journal of vocational training*, *37*, p.22.

Sabin, M., Alrumaih, H. and Impagliazzo, J., 2018, April. A competency-based approach toward curricular guidelines for information technology education. In *2018 IEEE Global Engineering Education Conference* (pp. 1214-1221). IEEE.

Vykopal, J. and Barták, M., 2016. On the design of security games: From frustrating to engaging learning. In *2016 {USENIX} Workshop on Advances in Security Education*.

Vykopal, J., Vizváry, M., Oslejsek, R., Celeda, P. and Tovarnak, D., 2017, October. Lessons learned from complex hands-on defence exercises in a cyber range. In *2017 IEEE Frontiers in Education Conference* (pp. 1-8). IEEE.

Weiss, R., Mache, J., Taylor, B., Kaza, S. and Chattopadhyay, A., 2019, February. Discussion of Integrating Hands-on Cybersecurity Exercises into the Curriculum in 2019. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (pp. 1245-1245). ACM.

Wiggins, G. and McTighe, J., 2005. Understanding by design. Alexandria VA: Association for Supervision and Curriculum Development.

Appendix 4

IV

S. Mäses, K. Kikerpill, K. Jüristo, and O. Maennel. Mixed methods research approach and experimental procedure for measuring human factors in cybersecurity using phishing simulations. In *ECRM 2019 18th European Conference on Research Methodology for Business and Management Studies*, pages 428–434. Academic Conferences and publishing limited, 2017

Mixed Methods Research Approach and Experimental Procedure for Measuring Human Factors in Cybersecurity Using Phishing Simulations

Sten Mäses[1], Kristjan Kikerpill[2], Kaspar Jüristo[3], Olaf Maennel[1]
[1]Department of Software Science, Tallinn University of Technology, Tallinn, Estonia
[2]Independent researcher, formerly at School of Law, University of Tartu, Tartu, Estonia
[3]Danske Bank
sten.mases@taltech.ee
kristjan.kikerpill@gmail.com
kaspar@cyberwiser.ee
olaf.maennel@taltech.ee

Abstract: Cyberattacks have a growing effect on business management. Organisations are increasingly focusing on human factors - how to train and evaluate people to minimise potential losses. One of the most scalable and practical ways to measure the human factor is to conduct a phishing experiment. Phishing is a type of cyber-attack that uses socially engineered messages to persuade humans to perform certain actions for the attacker's benefit. There is considerable amount of literature on the topic of phishing - e.g. how it works and how to fight against it. However, there is not much discussion on the particular methods nor the specific process of conducting simulated phishing experiments. This paper suggests a mixed methods approach for conducting phishing experiments and describes the experimental procedure including various technological, ethical and legal aspects. The suggested approach is based on related academic work and practical experience in both public and private sector organisations. Multiple opportunities and challenges regarding phishing experiments are discussed, providing guidelines for future research.

Keywords: mixed method research, cybersecurity, experimental procedure, phishing, human factor

1. Introduction

Phishing is a type of cyber-attack that uses socially engineered messages to persuade humans to perform certain actions for the attacker's benefit. It is a widespread and continuously evolving threat in cybersecurity forcing businesses to pay millions of dollars (Wardman 2016). To address the issue, security awareness campaigns and trainings are now being included into organisational security plans. For example, a reported 79% of companies in the US and 45% in the UK employ simulated phishing attacks to assess organisational susceptibility (Wombat Security, 2018).

Sending out a simulated phishing email might seem like a trivial task but should be carried out with much caution. There are multiple reports, where badly executed exercises lead to a situation where users do not click on any legitimate links anymore, or do not open any legitimate attachments at all anymore. This in turn leads to a loss of productivity and has a significant negative business impact. While multiple papers address different aspects of phishing, there is not much discussion on the particular methods nor the specific process of conducting simulated phishing experiments.

This paper describes a step-by-step process for carrying out a phishing campaign, starting from the general objectives and going through various legal, ethical and technical nuances. The legal context of European Union is used, where according to the General Data Protection Regulation (GDPR), non-compliance with personal data protection rules can entail severe business consequences for companies being subject to the regulations. A mixed methods approach is suggested for interpreting the results of the phishing experiment.

The described process is meant to be used as a guideline by any organisation that wants to carry out a phishing campaign, e.g. due to state imposed requirements on security or internal motivations for decreasing human factor vulnerabilities. The ideas presented in this paper have been collected and condensed from a series of academic papers, informal interviews with information security experts and experience from sending out around 500 simulated phishing emails targeting both private and public sector organisations, including many that are considered part of the critical information infrastructure.

The goal for this paper is not to provide a single, infallibly correct way for conducting phishing experiments, but to simulate a discussion within the community about the myriad of concerns and issues accompanying phishing experiments. We describe the phishing process from a viewpoint of a third party that is conducting a phishing test to an organisation. The process is very similar in the case where the organisation is conducting a phishing test internally, which is the preferred method of security evaluation, resources permitting.

2. Related Work

A considerable amount of literature has been published on the detection of phishing. Khonji and Iraqi and Jones (2013) have done a literature survey on the detection of phishing attacks and they provide a definition of phishing also used in the current paper: "Phishing is a type of computer attack that communicates socially engineered messages to humans via electronic communication channels in order to persuade them to perform certain actions for the attacker's benefit".

Dou et al. (2017) have systematically analysed different ways for software-based web phishing detection. Hadnagy and Fincher (2015) describe several aspects about phishing including the psychological principles behind phishing and how to recognise a phishing email. Jakobsson and Myers (2007) were among the first to comprehensively study phishing, providing a framework for studying the attack and its countermeasures. Their study describes how phishing works and what should be the defence mechanisms but lacks guidance on the process of phishing itself.

Attacks targeting the human factor get around various technical defence measures and have been the most popular methods employed by malicious actors in recent years (Wombat Security, 2018). The methods also enjoy a high success rate, with approximately 10% of phishing attacks successfully deceiving the recipient into clicking on a link or opening an infected attachment (Siadati et al, 2017).

At the same time, significant concerns remain regarding the frequency with which training and assessment tools are being employed. More than a half of the organisations only test susceptibility on a quarterly or yearly basis (Wombat Security, 2018). What is more, the most underestimated aspect of security related trainings is their actual impact on and efficacy in changing human behaviour (MacEwan, 2017). For this reason, Caputo et al. (2014) highlight the need to collect qualitative feedback from the participants after the phishing experiment, e.g. conduct interviews with the participants to gain a better understanding of how people behave in phishing experiments and why.

On the question of ethics in phishing experiments, Finn and Jakobsson (2007) concluded that when ethical aspects are not considered important or when neglected entirely, phishing simulation participants may get a sense of victimisation or irritation. Several other studies exist, which have found ways how to solve the ethical issues and measure users who are vulnerable for phishing attacks without causing them any distress (Jagatic et al, 2007). Likewise, Salah El-Din (2012) focuses on describing ethics committees' researchers' and professional bodies' perspective on ethical views about deceptive phishing research. Most importantly, it is outlined that the use of deception in phishing research can be safe, if done correctly.

3. Proposed process

In this chapter the general overview is given regarding the design and implementation process of a phishing campaign. We argue that it in most situations it is essential to use mixed methods approach instead of just measuring whether a person is tricked by simulated phishing email or not. Mixed methods approach means that qualitative methods are used for analysing content and human reaction. Also, quantitative measures are used to measure the person's reaction to phishing email. It is beneficial to measure not only whether a person was tricked by the phishing email but also how the person reacted in general. For example, it is good to know how many people notified the IT support or other point of contact appointed by the organisation.

There are multiple descriptions of phishing process from the criminal's perspective, but the guidance for conducting legal simulated phishing campaigns tends to be very general. We have identified 10 steps (Figure 1) necessary for designing and conducting a successful phishing experiment. Each step of the process is then analysed more in detail in the following sections.

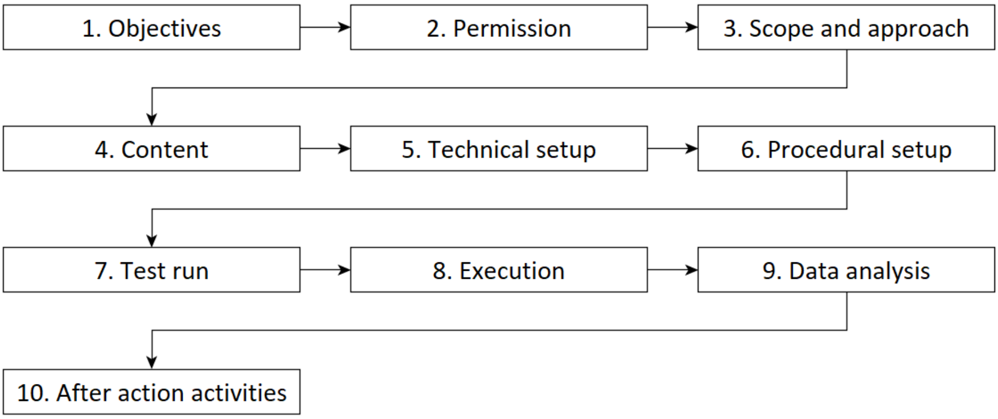


Figure 1 - General process for conducting a successful phishing experiment

3.1 Step 1 – Objectives

Defining the objectives for a phishing test is the starting point for the design of the phishing campaign. Given the gravity of issues related to phishing experiments, the necessity and the expected benefits of conducting phishing evaluations must be well-articulated in each specific case to avoid unnecessary irritation to employees, or "training fatigue" (MacEwan, 2017). Under usual circumstances, the objectives are mainly related to evaluating the human factor of cybersecurity in an organisation – as it concerns susceptibility to the actions of external malicious actors – but there could be multiple reasons behind it, e.g. the objective can be to understand the training needs for the personnel as well as the current level of knowledge and experience of employees. Furthermore, the objective could also be to estimate the effectiveness of specific security training provided to the personnel previously or to comply with regulations (Hadnagy and Fincher, 2015).

As a rule, it is not recommended to conduct any phishing tests before every targeted employee has gone through security training or instruction. The only exception in this case would be doing the phishing test as an introduction just before a planned security training, to make things more "real" for the employees in subsequent instruction, and as a baseline for assessing the effectiveness of subsequent security training. It is of paramount importance to avoid any blaming or shaming based on the results of the test. Many researchers have shown that fear is an ineffective tactic to motivate security-related behavioural change (Bada and Sasse and Nurse, 2015). People should not be punished based on the experiment, but rather given a detailed explanation, and/or the opportunity to attend extra training or read additional materials about the topic of secure behaviour.

An important aspect to note when considering the creation, changing or amendment of organisational security policies, is that requests for compliance are more likely to be followed if they are perceived as fair, consistent and legitimate (Wortley, 2013). It is therefore necessary that figures of authority within an organisation make sure all such changes are promptly and adequately communicated to the members of the organisation. Only then would subsequent qualitative feedback benefit further amendments or adjustments to internal policy rules.

3.2 Step 2 – Permission

This section discusses several legal requirements for conducting a lawful phishing experiment. Before researchers or security testers can perform a phishing experiment, it is crucial to obtain proper written permission from the management of the targeted organisation.

Obtaining proper, and sufficient, permissions cannot be overemphasised regarding any types of phishing experiments or evaluations, as these have deception at their core. Permissions are the thin line between a well-intended phishing evaluation beneficial for organisational security and a malicious, or unidentified, "in-the-wild" phishing campaign. From the perspective of penetration testers, security auditors or other type of researchers external to the organisation, no-permission phishing is usually illegal. Even if the activity would not be subject to criminal liability, then it is without a doubt subject to personal data protection safeguards, since data on specific email accounts will be collected in each case. A proper permission means a written and signed document obtained from the person or body, e.g. the board, invested with the according authority. If the chief technology

officer or head of security has been authorised to make such decisions personally, then not only this permission should be sought, but also a clarification letter that expresses the limits of his/her responsibility.
Signing a non-disclosure agreement with the specific company is crucial, allowing for a quid pro quo approach that serves both research interests and provides useful information to the organisation about their security needs. In general, this step of the process will also shed light on how the security related decision-making is operationalised within the organisation, e.g. how much freedom is given to the security personnel to conduct necessary vulnerability testing as well as what permissions are needed and whether legal counsel is available. It is a common practice to have written permission and non-disclosure agreement between the organisation and the external party providing the phishing simulation. However, there is anecdotal evidence showing that having official written permission is less common for an internally organised phishing test.

3.3 Step 3 – Scope and approach

In the step 3, the target audience for the experiment should be determined. Also, a general approach and timeline should be set.

3.3.1 Selecting the sample group

Once the larger objectives have been set and permissions granted, then it is time to define the scope of the experiment. Based on the objectives, a reasonable scope should be determined. Usually, it is not feasible to send the phishing email to every member of the organisation. Also, it is meaningless to send the phishing email to people who are on holidays or otherwise not expected to read their email.
In some organisations, it makes sense to divide people into groups based on their job description. There could be different target groups such as management, computer specialists etc. It is better to choose those people who are trained and informed that a phishing experiment may take place. Also, in case of international organisations, legal jurisdictions should be considered. For example, a written permission to perform a phishing experiment in a local branch of the organisation might not cover some employees that are officially part of another branch of the organisation or otherwise hired under different conditions.

3.3.2 Timeline and general process

Phishing could be done as a one-time experiment or a part of an ongoing series. The sample group can be divided into two or more subgroups to receive phishing email with different content (crossover trial). That approach enables to better understand the impact of the phishing email content to the user's reaction. The crossover trial approach is illustrated in Figure 2.

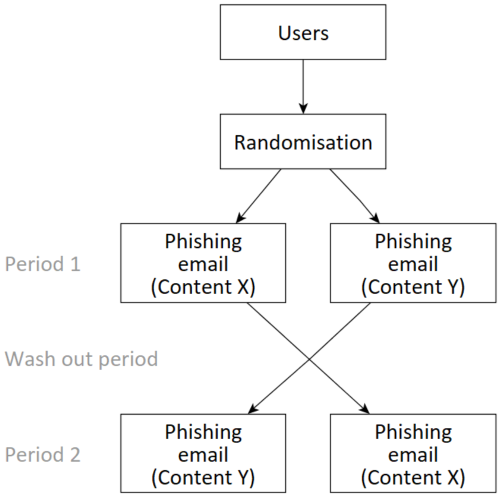

Users
Randomisation
Period 1
Phishing email (Content X)
Phishing email (Content Y)
Wash out period
Period 2
Phishing email (Content Y)
Phishing email (Content X)

Figure 2 - Crossover design for phishing email content to better understand the impact of different types of content

3.3.3 Selecting phishing type

There are numerous types of phishing attacks. For example, Chiew and Yong and Tan (2018) have mentioned 20 different types. The main actions that the targeted person is expected to perform are the following:

- visit a website (e.g., to be infected by malware or to enter personal information);
- reply to the email (e.g., to provide personal or otherwise sensitive information);
- open the attached file (e.g., to be infected by malware);
- transfer money (e.g., react to a fake invoice or blackmail scam).

When simulating a phishing attack, a suitable attack vector must be selected. Usually it is legally questionable and impractical to ask people to transfer money. Based on our experience, people are not easy to convince to give up sensitive information by email reply either.

Sending an attached file by email and measuring the user behaviour can be quite realistic simulation but also technically challenging. It is difficult to create a simulated malware that would pass any potential anti-virus checks, work on different operating systems (Linux, Windows, macOS, etc), and be legal to distribute. If the custom-made email attachments fail for any technical reason (not due secure behaviour decision of the employee), then the reported opening rate could be seriously biased.

Inviting targeted users to visit a website is a convenient attack method for measuring the response. Nevertheless, it also contains multiple design choices. The hyperlink in the phishing email could lead to either a so-called meaningless site (website is blank, redirects to some other site, displays an error or an infinite loading message), seemingly legitimate site (an existing company or a fictional site) or information about the experiment together with optional guidance for further action (e.g., security policy or training materials). Copying a legitimate website (such as Gmail login form) is strongly discouraged without the explicit permission from the site owner. Showing information about the experiment is a good way to avoid potential confusion and overreacting but at the same time it does not allow to measure how many people would report the phishing site after clicking the link.

Selection of the phishing type influences the content of that phishing email. More about that in the next section.

3.4 Step 4 – Content

The main requirements for the content of the phishing email are for it to be not offensive (no threats, insults etc) and to contain multiple suspicious characteristics that would allow the receiver to decide that it is a phishing email.

Hadnagy and Fincher (2015) have described different difficulty levels for phishing emails. In our experiments, we used two different email contents for each experiment (Figure 2). One content was more suspicious and the other was less suspicious and therefore more difficult to recognise as phishing. If difficulty level of the phishing email is very high (almost authentic email), then it is difficult for ordinary users to understand what they did wrong and as a result they might become overly cautious (not clicking on links in legitimate emails) or experience security “fatigue” and decrease their focus on security behaviour (MacEwan, 2017). Characteristic mistakes in the phishing email enables to better educate users afterwards, showing them the specific signs that the received email was fraudulent.

3.5 Step 5 – Technical setup

The technical setup is largely influenced on the selected phishing type (step 3). It covers technical design choices for optimally sending out phishing emails and measuring the click rate.

One of the first technical questions that needs an answer is whether to buy the appropriate service or to set up an inhouse environment for sending phishing email campaigns. It is assumed that the phishing campaign will be conducted legally, therefore intentional use of illegal or questionably legal services, such as Phishing as a Service platforms in the dark web (Li et al, 2013), is out of the scope. There are many organisations that provide legal phishing as a service, e.g. Nexigen, KnowBe4, Guardian360, Cofense to name a few. (Authors of this paper do not have any affiliation with any of the companies mentioned.) Using an external service to manage the email sending part or the whole phishing process helps to simplify things for the end user and lessen the administrative overhead. The main downside is the lack of control over the experiment and the received data. The current paper is mostly targeted to those situations where the phishing as a service is not used.

Fincher and Hadnagy (2015) have compared different software platforms for phishing experiments. There are various open-source and commercial platforms available. In our case, we have used open-source solutions. Available open-source solutions include, e.g., Gophish, King Phisher, and Phishing Frenzy. They generally require some effort to get the system up and running but provide more granular control over the process. In our case study the deciding factor was privacy. In our experience, none of the targeted organisations did agree to take any chances for leaking their sensitive data (employee email addresses and phishing campaign results). Therefore, we decided to choose one of the open source platforms and run it in our own server (belonging to the university) and special attention was put to storing the gathered information securely (e.g., disconnecting the phishing server from the Internet after the data gathering is complete).

To minimise any potential technical issues on measuring the click rate, additional data measurement points can be used (e.g., storing the network traffic with tcpdump).

3.6 Step 6 – Procedural setup

Procedural setup phase focuses on informing all the concerned parties. In our case we used university server for sending out emails and hosting the landing page. We also used domain name registrar to redirect the custom web address to our university server. Therefore, we had to inform the IT department of university (people responsible for incident handling and security monitoring) and the domain name registrar about the planned phishing experiment. We also informed the national computer emergency response team (CERT) and the point of contact in the targeted organisation.

Keeping relevant people informed avoids potential overreaction. For example, CERT can rest assured that it is a simulated attack, not an actual emergency. In our case, the mentioned parties got precise information about the planned time for the phishing experiment and the exact content of the phishing emails.

Additionally, the point of contact in the targeted organisation kept track on all the relevant reports by the employees noticing a suspicious email. As also emphasised by Hadnagy and Fincher (2015), it is essential to gather data on incident reporting.

Another discussion point is whether to notify employees beforehand or not. Fully notifying people beforehand causes bias in the results, since people tend to be more attentive, if they possess specific prior warning as well as knowledge of ongoing monitoring regarding their activities and reactions. However, the opposite solution of complete deception is wrought with ethical, and potentially legal problems. Based on our observations, a prior warning multiple weeks (or even months) in advance does not cause significant bias.

3.7 Step 7 – Test run

No matter, how good is the plan, there is often some technical detail that is overlooked. That's why it is essential to conduct a test run before the actual phishing campaign starts. We recommend conducting the test run at least one day in advance, so that there is still time to fix some minor technical issues or to postpone the main campaign in case of bigger difficulties. Test run should ideally be conducted using an actual email address of one informed employee of the organisation.

3.8 Step 8 – Execution

After careful planning and testing, it is time to execute the campaign. In our case, it included sending additional notifications to national CERT and target organisation just before the start of the experiment. Then we sent out the phishing emails with two different types of content according to the crossover trial approach (Figure 2). Emails were sent usually around 1 PM and contained personal link to a website showing an error "Page not found". Around 4 PM another email was sent to targeted employees by the information security responsible in that organisation. That email explained that a phishing experiment took place and provided additional information about correct security behaviour.

3.9 Step 9 – Data gathering and analysis

Visits to the website mentioned in the phishing email were recorded. As targeted people were given personalised links (in the format of www.example.com/?c28da2, where "c28da2" was personal identifier), then people clicking on those links could be determined. Some people did not use the personalised link, but instead the main website (www.example.com). Their visits were logged but could not be connected with their identity. Additionally, all the reports to the targeted organisation's security team were logged and later analysed to see whether the reporting person had already clicked on the phishing link or not. As a general trend, we observed around 10% click rate related to simpler (more obviously fake) emails and around 20% click rate with the more sophisticated (less obviously fake) emails. Although those findings are well aligned with other phishing research, e.g., Siadati et al. (2017), it must be understood that some special cases might not be interpreted correctly. For example, it could be argued that if an advanced computer user uses proper security measures (e.g., dedicated sandbox computer with activated security plugins such as NoScript), then just visiting a website is not very risky. Nevertheless, it would be expected that this advanced computer user reports the suspicious email.

As for the reporting, we have observed massive underreporting. Around 75% of the people who did not click on the phishing link also did not report it. That shows that there is a big chance of a more targeted phishing attack to go unnoticed by the security team even if some targeted people understand that they have been targeted. Therefore, it is important to measure reporting to get an overall understanding about how people behave after getting a suspicious email.

Interviewing some phishing experiment targets helps to interpret the results and is highly recommended. We selected interview participants based on their reaction (clicking on the phishing link and reporting suspicious email). Quasi-structured interviews explained, for example, why some phishing content was less successful (unknown sender, unrelated to work) and why there was a lack of reporting (not finding it urgent, not knowing reporting processes). Interviewed people confirmed the need of keeping the content of the simulated phishing emails not threatening nor offensive. This resonates well with the suggestions by Finn and Jakobsson (2007).

3.10 Step 10 – After action activities

The last step of the phishing exercise is to tie up any loose ends. The main focus in this phase is usually on delivering a report covering the results and analysis of the experiment. It would be good to also emphasise in the report that the targeted people should not be punished in any way. The report should provide a good basis for the organisation management or other decision makers to take the next steps to improve their security levels.

A careful debrief with the targeted employees helps them better understand why this experiment was needed and avoids damaging the trust between them and the management. The exercise should be treated as an input for considering the general training needs of the personnel. While victimisation is hopefully avoided, there should still be enough resources ready to handle potential misunderstandings. People can understand emails differently and react in various unexpected ways. In the worst-case scenario, even measures such as psychological counselling might have to be considered. Although such extreme reactions might be rare, it still is possible that several employees are left confused after the experiment and require further information (in addition to the explanatory email that is sent in the end of the phishing experiment). Therefore, it would be a good idea to offer employees an opportunity for an individual debrief explaining the aims of the experiment and addressing any potential misunderstandings or concerns they might have.

In the end, it is recommended, and in many contexts legally required, to securely delete any sensitive data that was gathered during the exercise or save the data in an anonymised form.

4. Conclusions

Conducting a phishing campaign is a relatively easy and well scalable way to get some insights into the current security posture of an organisation. While the process of sending an email to the employees might seem quite straightforward at first, the process has numerous details that are important to notice, but easy to miss. Overlooking some of those design choices, e.g. the legal implications and policy environment, might severely backslash the good intention.

Also, the analysis of the results should be done with caution. This paper described how to use a mixed method approach where after sending simulated phishing emails to employees, quasi-structured interviews were held with a selection of targeted people. Qualitative side of the experiment helps to better interpret the quantitative results. This paper is intended to initiate a more in-depth academic discussion within the community on how to conduct awareness trainings that are positive and address the real threats.

Acknowledgments

We would like to thank all the organisations that agreed to take part in phishing experiments and CERT-EE for their cooperation. Also, we would like to acknowledge the European Regional Development Fund of European Union for supporting this research.

References

Bada, M., Sasse, A.M. and Nurse, J.R., 2015. Cyber security awareness campaigns: Why do they fail to change behaviour?. International Conference on Cyber Security for Sustainable Society.

Caputo, D.D., Pfleeger, S.L., Freeman, J.D. and Johnson, M.E., 2014. Going spear phishing: Exploring embedded training and awareness. IEEE Security & Privacy, 12(1), pp.28-38.

Chiew, K.L., Yong, K.S.C. and Tan, C.L., 2018. A survey of phishing attacks: their types, vectors and technical approaches. Expert Systems with Applications.

Cofense PhishMe, [online], https://cofense.com/product-services/phishme/

Dou, Z., Khalil, I., Khreishah, A., Al-Fuqaha, A. and Guizani, M., 2017. Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection. IEEE Communications Surveys & Tutorials, 19(4), pp.2797-2819.

Finn, P. and Jakobsson, M., 2007. Designing ethical phishing experiments. IEEE Technology and Society Magazine, 26(1), pp.46-58.

GDPR, 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Gophish, [online], https://getgophish.com/

Grazioli, S., 2004. Where did they go wrong? An analysis of the failure of knowledgeable internet consumers to detect deception over the internet. Group Decision and Negotiation, 13(2), pp.149-172.

Guardian360, [online] https://www.guardian360.net/solutions/phishing-as-a-service/

Hadnagy, C. and Fincher, M., 2015. Phishing Dark Waters: The Offensive and Defensive Sides of Malicious Emails. John Wiley & Sons.

Jagatic, T.N., Johnson, N.A., Jakobsson, M. and Menczer, F., 2007. Social phishing. Communications of the ACM, 50(10), pp.94-100.

Jakobsson, M. & Myers, S., 2007. Phishing and countermeasures: understanding the increasing problem of electronic identity theft, Hoboken, NJ: Wiley-Interscience.

Khonji, M., Iraqi, Y. and Jones, A., 2013. Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials, 15(4), pp.2091-2121.

King Phisher, [online], https://github.com/securestate/king-phisher

KnowBe4, [online], https://www.knowbe4.com/security-awareness-training-features/

Li, Z., Alrwais, S., Xie, Y., Yu, F. and Wang, X., 2013, May. Finding the linchpins of the dark web: a study on topologically dedicated hosts on malicious web infrastructures. In Security and Privacy (SP), 2013 IEEE Symposium on (pp. 112-126). IEEE.

MacEwan, N., 2017. Responsibilisation, rules and rule-following concerning Cyber Security: Findings from Small Business Case Studies in the UK (Doctoral dissertation, University of Southampton).

Nexigen, [online], https://nexigen.com/it-services/phaas/

Phishing Frenzy, [online], https://www.phishingfrenzy.com/

Salah El-Din, R., 2012. To deceive or not to deceive! Ethical questions in phishing research.

Siadati, H., Palka, S., Siegel, A. and McCoy, D., 2017, August. Measuring the Effectiveness of Embedded Phishing Exercises. In 10th {USENIX} Workshop on Cyber Security Experimentation and Test ({CSET} 17). USENIX} Association}.

Soghoian, C., 2008. Legal Risks for Phishing Researchers. The Third Anti-Phishing Working Group eCrime Researchers Summit.

Wardman, B., 2016. Assessing the Gap: Measure the Impact of Phishing on an Organization.

Wombat Security (2018) “State of the Phish 2018”, [online], https://info.wombatsecurity.com/stateof-the-phish

Wortley, R., 2013. Situational precipitators of crime. In Environmental criminology and crime analysis (pp. 70-91). Willan.

Appendix 5

V

S. Yari, S. Mäses, and O. Maennel. A method for teaching open source intelligence (OSINT) using personalised cloud-based exercises. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security: ICCWS 2020*. Academic Conferences and publishing limited, 2020

A Method for Teaching Open Source Intelligence (OSINT) Using Personalised Cloud-based Exercises

Saber Yari, Sten Mäses, Olaf Maennel
Tallinn University of Technology (TalTech), Tallinn, Estonia
saberyari@protonmail.com
sten.mases@taltech.ee
olaf.maennel@taltech.ee

Abstract: The ability to analyse publicly available information is highly valued in the cybersecurity community and is often crucial in areas such as incident investigation and penetration testing. However, it has been challenging to provide practical hands-on exercises to train the techniques relevant to open source intelligence (OSINT). This is particularly so as gathering sensitive data about a real person or organisation in a classroom exercise is likely to bring out a multitude of ethical and legal issues.

To conduct classroom-based OSINT exercises it is possible to create a sandbox environment with simulated services or to create some publicly available services that can be analysed by the students. However, both approaches having their shortcomings. In a sandbox environment, it is very difficult to simulate relevant real-world systems such as Facebook, Google, or Shodan. Setting up public services to be available to everyone is also problematic as it is usually not possible to provide an individual student experience. Therefore, once one participant has found the answer, it is very easy to cheat and share the solution with others.

This paper presents a novel way for constructing a learning environment focusing on OSINT exercises where capture the flag (CTF) style tasks are conducted in cloud based virtual labs. The students use real applications throughout the exercise environment with the network traffic in the virtual lab routed through a specialised proxy. Here the contents of the web sites are modified to contain flags that are individual for each participant. Assessment approaches and scenario development are discussed together with key learning points from conducting this case study. This experimental setup is already being used in a university undergraduate level introductory course in cybersecurity at the Tallinn University of Technology.

Keywords: cybersecurity education, gamification, cybersecurity exercises, OSINT

1. Introduction

The term OSINT originates from military usage and is defined here as the gathering and processing of publicly available information. Together with SIGINT (Signals Intelligence), which is intelligence gained from signal intercepts and HUMINT (information from human sources), OSINT is regarded as one of the three traditional intelligence sources (Best, 2008). As the amount of information that is publicly available from the Internet grows, finding and analysing relevant material is a highly valued skill. However, despite this there is a lack of research done on the topic of training OSINT collection techniques. This paper focuses on OSINT from a cybersecurity context.

We propose that the lack of hands-on trainings for OSINT-related skills is due to inherent technical difficulties in creating realistic training environments while providing a personal learning experience. While cybersecurity exercises are increasingly used to train specialists to find and patch vulnerabilities from computer systems, these exercises rarely cover OSINT-related tasks and skills. Many cybersecurity exercises utilise capture-the-flag (CTF) formats where participants are challenged to find or protect specific information (flags). While it is possible to generate individual flags for each participant in a normal cybersecurity exercise context, it is much more difficult to do it for OSINT tasks. For example, an OSINT task may include geo-locating an image posted on social media by examining the metadata generated when it was created. However, it is difficult to ensure that after the first student has discovered this information that it is not then shared among their classmates. For that reason, OSINT competitions like the TrendMicro OSINT Challenge[1] are generally not suitable for a classroom education context. An alternative method may be to create separate social media profiles with unique material and custom metadata for each exercise participant. However, this approach is not usually feasible because it does not scale well to large numbers of students. A compromise solution may be to simulate public web sites such as Google and Twitter in the exercise environment. However, this too is problematic as popular web sites tend to be very complex and data intensive. Additionally, the learning experience from engaging with an artificial, simulated public web site such as Google search may be very different from the experience of engaging with the live Internet.

[1] https://resources.infosecinstitute.com/trend-micro-osint-challenge/

This paper presents an alternative method overcoming the challenges of conducting capture-the-flag (CTF) style OSINT training using cloud based virtual labs. This enables students to access real applications through the virtual lab environment. However, as the network traffic in the virtual lab is routed through a specialised proxy, the content of the web sites is modified to contain flags that are unique to each participant. This experimental setup is currently being used in an introductory course to cybersecurity at the Tallinn University of Technology. Assessment approaches and scenario development are discussed together with key learning points from conducting this case study.

2. Background and related work

The systematic gathering of OSINT arguably originates from the Second World War when the Foreign Broadcast Monitoring Service was established by the US (Schaurer and Störger, 2013). The term OSINT itself was coined in the late 1980s by the US military, signifying a greater emphasis on freely available, rather than secret information sources (Schaurer and Störger, 2013). The expansion of Internet technologies has rapidly increased the potential of OSINT. It has been estimated that OSINT can provide between 80% and 98% of all information that a government or private sector organisation needs to know for making informed decisions (Gibson, 2014). Although these figures are difficult to precisely measure and depend on the circumstances, they do emphasise the importance and high value of OSINT sourced data.

There are several papers discussing the role of OSINT (e.g., Gibson, 2014) and its use in different contexts. Wells and Gibson (2017) compare the use of OSINT in law enforcement and military domains. Quick and Choo (2018) present a framework for supporting digital forensic investigators by using OSINT methods for locating relevant information, which provide context and added value to previously collected data. However, despite these previous studies, there is a lack of research on the educational side of how to teach OSINT collection techniques. This is in stark comparison to cybersecurity, where there is a considerable amount of literature examining both the general concepts of education as well as focusing on specific topics. Rege (2015) for example discusses the weaknesses of CTF exercises but concludes that the benefits far outweigh the shortcomings. Furthermore, she argues that the CTF type of multidisciplinary cybersecurity exercises can provide valuable learning experiences for both computer science and criminology students. Konak (2018) presents a weeklong program to teach K-12 students cybersecurity concepts and skills through hands-on activities. He shows how practical cybersecurity exercises based on pedagogical approaches can improve the self-efficacy of the participants. Mäses et al (2018) also describe a framework for turning the results of cybersecurity exercises into a more meaningful evaluation of each participant's individual skills.

Other published research focuses specifically on practical cybersecurity training. For example, Yuan et al (2017) evaluate a variety of hands-on labs for teaching SQL injection and Figueroa et al (2018) present a method for teaching cybersecurity aspects of Radio Frequency Identification. The value of cybersecurity exercises to simulate cyberattacks and defence dynamics has been shown (Caliskan et al, 2017), but the systems that are employed are usually separated from the public Internet.

Gamification is another means being used to increase engagement by teaching through serious games or virtual simulations (Hamari et al, 2014). Morschheuser et al (2017) approaches gamification via a design science research approach and suggest a method for its configuration. Chou (2015) discusses how to make gamification actionable and presents the Octalysis framework to provide a systematic approach to core motivating factors that are implemented in gamification.

Together, these show that hands-on cybersecurity exercises are highly valued in the scientific community. However, although it makes sense to incorporate OSINT-related tasks to the cybersecurity exercises there is little publicly available research investigating the use of virtual labs to develop OSINT-related skills. So far, this is because the OSINT skills that would be useful in real situations cannot be fully utilised in the exercise platforms due to the design constraints. This paper aims to address this shortcoming by introducing a method for teaching OSINT using personalised cloud-based exercises.

3. Methodology

The goal of this study is to demonstrate how OSINT-related tasks can be incorporated into a virtual lab environment. To test the suggested method for creating personalised hands-on tasks related to OSINT, the case study method was chosen. The resulting OSINT virtual lab was trialled in an undergraduate level cybersecurity course. A crossover design was utilised in distributing the homework to gain further insights into the influence of timing of the assignment. Various gamification elements were also used to improve the engagement of the participants. After completing the given OSINT assignments, feedback was collected using a questionnaire. The following sections describe the design process, including considerations for the lab content, addition of gamification elements, and the technical innovations used. These enabled the lab to be both realistic and individual at the same time.

3.1 Technical design

Developmental research was used to construct a virtual lab to enable OSINT training to be conducted in a personalised way. Developmental research was used to answer questions about how different artefacts could be constructed to address the problems and case studies produced to prove that the suggested solutions were viable. In developmental research, after establishing and validating product criteria, a process for product development is accepted and formalised with the finished product evaluated based on set criteria (Ellis and Levy, 2009). Developmental research is widely used in software development and was considered appropriate to the aspects of this current research investigating technical software development.

The main requirements for the technical design were as follows. *First*, the solution must be scalable and so it was decided that it would be implemented in a cloud-based environment with the scoring conducted automatically. *Second*, for every participant, a different answer must be generated to prevent possible cheating attempts. *Third*, the user experience must be realistic – that is, the skills acquired from the gamified virtual lab must be easily transferrable to a real-life environment. To meet the first and second requirements, an open source virtualisation platform i-tee (Ernits and Kikkas, 2016) was used. This platform enables participants to remotely access the cloud-based virtual labs using their web browser. At the same time, it enables automated scripts to use the participant's username to create personalised flags.

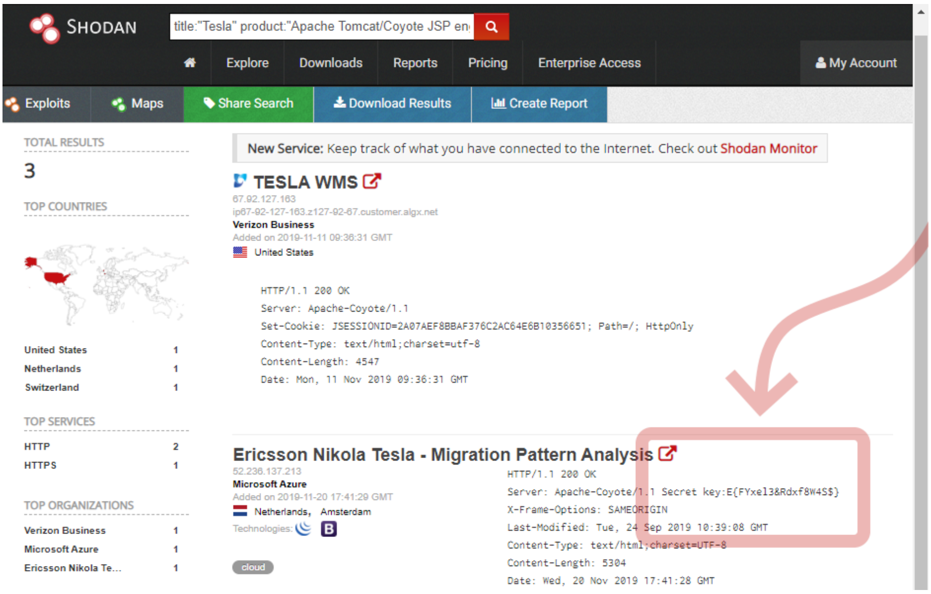
Although the i-tee platform was used, the method for creating realistic personalised OSINT is not dependent on the specific underlying virtualisation platform, but rather consists of several techniques. These include a self-generated certificate authority, reverse proxy, Domain Name System (DNS) cache poisoning, redirecting Hypertext Transfer Protocol (HTTP) and Hypertext Transfer Protocol Secure (HTTPS) traffic. These techniques together enable the end-to-end encryption that most websites use to be broken and personalised flags inserted to the network traffic (Fortinet Inc, 2017). There are multiple ways of breaking HTTPS encryption including DNS cache poisoning and redirecting traffic through a sink-hole, both of which are discussed below. In our research, both DNS cache poisoning and the sink-hole technique to redirect network traffic were used to insert flags. During these exercises, participants should be, and in our case were, briefed not to access private information when using the environment due to network traffic encryption being purposefully broken. However, by using a virtual lab environment, participants were well aware that they were operating in an artificial exercise context. At the same time, having a game-environment that breaks encryption enabled the third requirement or creating a realistic user experience to be met. The techniques used are described in detail in the following sections with additional technical sample code given in the appendices.

3.1.1 Self-generated certificate authority

Certificate Authorities (CA) issue Secure Sockets Layer (SSL) certificates for associating ownership information with cryptographic keys. Internet browsers use these certificates to verify the website's identity. Self-generating CA is a common practice for corporations and anti-virus software to monitor the activities of the targeted user. Installing a fake CA is also a known method used by cybercriminals for performing man-in-the-middle attacks. (Cui et al, 2018; Durumeric et al, 2013). Self-generated certificates allowed us to issue fake SSL certificates for the chosen public websites and avoided a warning message of insecure communications being shown to users. This contributed to making their user experience smoother and more realistic.

3.1.2 Reverse proxy

Reverse proxy is a type of proxy server used in load balancing, cache optimization and intrusion detection by enabling resources to be retrieved on the client's behalf (Stricek, 2002). Reverse proxy can also be used for malicious purposes to hide a back-end server from other users (Yoshida, 2008). Implementing a reverse proxy in the exercise lab allowed us to retrieve the HTTP responses that have been redirected via DNS cache poisoning. Reverse proxies can also modify the received content (Stricek, 2002). This feature allowed us to alter the responses and insert personalised flags. Appendix 1 shows the configuration used for this technique and Figure 1 illustrates how individual flags can be inserted to specific publicly available web sites.



TOTAL RESULTS

3

TOP COUNTRIES

United States	1
Netherlands	1
Switzerland	1

TOP SERVICES

HTTP	2
HTTPS	1

TOP ORGANIZATIONS

Verizon Business	1
Microsoft Azure	1
Ericsson Nikola Te...	1

New Service: Keep track of what you have connected to the Internet. Check out Shodan Monitor

TESLA WMS

67.92.127.163
ip67-92-127-163.z127-92-67.customer.algx.net
Verizon Business
Added on 2019-11-11 09:36:31 GMT
United States

```
HTTP/1.1 200 OK
Server: Apache-Coyote/1.1
Set-Cookie: JSESSIONID=2A07AEF8BBAF376C2AC64E6B10356651; Path=/; HttpOnly
Content-Type: text/html;charset=utf-8
Content-Length: 4547
Date: Mon, 11 Nov 2019 09:36:31 GMT
```

Ericsson Nikola Tesla - Migration Pattern Analysis

52.236.137.213
Microsoft Azure
Added on 2019-11-20 17:41:29 GMT
Netherlands, Amsterdam
Technologies:
cloud

```
HTTP/1.1 200 OK
Server: Apache-Coyote/1.1 Secret key:E{FYxel3&Rdxf8W4S$}
X-Frame-Options: SAMEORIGIN
Last-Modified: Tue, 24 Sep 2019 10:39:08 GMT
Content-Type: text/html;charset=UTF-8
Content-Length: 5304
Date: Wed, 20 Nov 2019 17:41:28 GMT
```

Figure 1: Individual flag inserted into Shodan search engine result

Flags consisted of sixteen random characters and were automatically generated after each lab initiation. Flags were typically included in the HTTP response using a man-in-the-middle technique. Users investigating the relevant web sites outside of the lab environment would only see a normal response without any flag or a random string not giving them any useful information. Reverse Proxy replaced these random strings with the real flags. With this technique, it was possible to ensure that each user receive a different flag. They could also only find the flags if they accessed the target web address on their own lab instance. Appendix 2 shows the script used to attain this goal.

3.1.3 *DNS cache poisoning*

Domain Name System (DNS) cache poisoning or DNS spoofing is one of the most prominent DNS attacks. A DNS resolver stores an invalid record for the DNS queries, which causes the domain to be mapped to an alternative IP address (Son and Shmatikov, 2010). The virtual machines of the exercise labs were in full control of the lab developer. This meant that it was possible to deliberately resolve the DNS queries of users to an internal server instead of their original IP address. An example script illustrating how the implementation of this technique was created in the OSINT lab is shown in Appendix 3. Figure 2 illustrates a summary of this technique.

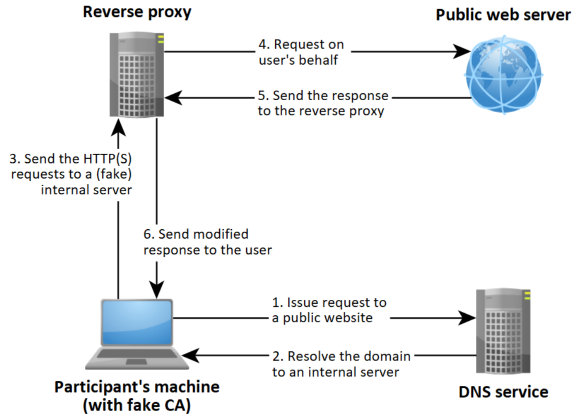

Reverse proxy
Public web server
4. Request on user's behalf
5. Send the response to the reverse proxy
3. Send the HTTP(S) requests to a (fake) internal server
6. Send modified response to the user
1. Issue request to a public website
2. Resolve the domain to an internal server
Participant's machine (with fake CA)
DNS service

Figure 2: Modifying user traffic using DNS Cache Poisoning

3.1.4 *Redirecting HTTP and HTTPS traffic*

When DNS poisoning is used to map the domains to an internal IP address, the user's machine receives an alternative IP address instead of the real address. If users investigate the IP address accessed, DNS poisoning prevents them from accessing the genuine mapped IP address of the domain and prevents its further examination. To address this limitation, iptables are used to change the destination of packets and eventually redirect HTTP and HTTP traffic to a different IP address (Kaur and Elsner, 2015). Redirecting users without DNS poisoning allows users to see the real IP address of the destination via regular DNS queries, but the lab server redirects the traffic to an internal web server without notifying the users. Figure 3 shows a diagram illustrating how modifying user traffic using iptables works.

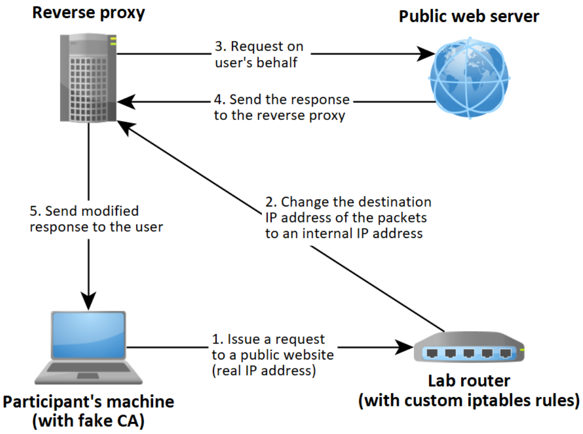

Reverse proxy
Public web server
3. Request on user's behalf
4. Send the response to the reverse proxy
5. Send modified response to the user
2. Change the destination IP address of the packets to an internal IP address
1. Issue a request to a public website (real IP address)
Lab router (with custom iptables rules)
Participant's machine (with fake CA)

Figure 3: Redirecting and modifying user traffic using iptables

3.2 Content design

The topics and tools that were chosen for the exercise were inspired by the framework proposed by Quick and Choo (2018) and from the list of tools in the OSINT Framework[2]. The OSINT-related learning topics and tools that were selected for the lab were wide ranging. They included Google Advanced Search[3], Shodan[4], Wayback Machine[5], analysis of meta-data of files such as EXIF data in picture files, and geo-location through GPS and web maps. In addition, person and email search using specialised search engines as well as breached email databases, social media search, analysis of email headers and the source code of web sites were used.
To improve the engagement, several gamification elements were used from Chou (2015), focusing on white hat engagement core drivers of the Octalysis framework (Chou, 2015). Some examples include having a scoreboard, Easter eggs (secret messages not directly connected to the score of the game), and a heroic storyline. The narrative was constructed as a mission of finding a hacker that has tricked a victim using a phishing email. The students had to first analyse the email source and then continue the investigation using the mail server IP address as the starting point. Using OSINT techniques, it was possible to obtain the (fictional) attacker's name, picture and location. For ethical and legal considerations, only fictional characters were used in the storyline. Fake Name Generator[6] was used for naming the fictional personas. Morphthing[7] was used for creating profile photos by merging together multiple photos that were licenced for free reuse with modifications.

[2] https://github.com/lockfale/OSINT-Framework
[3] https://www.google.com/advanced_search
[4] https://www.shodan.io/
[5] https://archive.org/web/
[6] https://www.fakenamegenerator.com/
[7] http://www.morphthing.com/

4. Implementation

In the initial implementation phase, twenty-six different flags were created for the users to find. These flags were inserted in pictures, app profiles, emails, 10 public web sites, and 4 in-house developed websites to simulate the sites that are owned by the hacker. Since flags are individual strings, it is easy to remove or add new flags to the game in a modular manner. This makes it possible to change the location of the flags or customize them more efficiently. Five master (graduate) level students not participating in the main course were asked to test the initial lab and give their feedback. Based on the feedback, the lab guides were revised to be easier to understand and more specific and a dynamic graph visualising the discovered and undiscovered flags was added. The OSINT lab was then used by 94 students enrolled on a cybersecurity course. Students were divided into two groups – both having one week to complete the lab. However, one group was asked to participate immediately after an introductory lecture and the other one week later.

5. Results and discussion

Overall there were 26 flags in the OSINT lab. The number of flags found by the participating students was between 1 and 22. The number of flags found per users is shown in Figure 4.

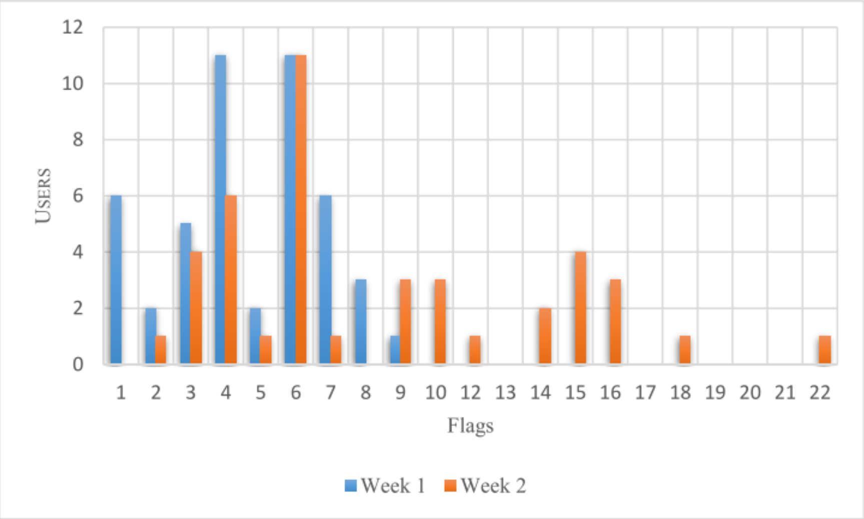


Figure 4: Flags found by users during week 1 (group 1) and week 2 (group 2).

Figure 4 clearly shows that students in the second group were able to find many more flags than the first group. One possible reason could be that students were being given hints and the location of the flags by the earlier group. Another interesting observation is that in the first day of the exercise, students of the second week were able to find four times the number of flags found in the first day by the first week's group. It took more than three days for the first group to find the same number of flags that the second group found during the first day. This raised the question whether the sudden increase at the beginning of the second group's exercise was because some students shared how to find the flags with other students. Another observation concerns the flag that was found by sending a real email to an address that could only be found during the exercise. It took eight days after the lab was initially given to students for the first email to be sent. However, after the flag was successfully retrieved six different correct emails rapidly followed providing another indicator of how fast the methods for receiving flags were conveyed amongst the students.

To further investigate the information sharing aspect, an additional e-mail was sent out to students after the end of semester asking for their honest view on the homework solving process. 6 students replied. Out of those, 3 students said that they did the exercise independently and did not notice any information sharing. Another 3 students noticed some information sharing – mostly generic ideas shared in smaller groups. One of the students who noticed information sharing mentioned that the information offered was not very helpful. Another student mentioned that the difference in results might have been caused by varied study load in different weeks. Two cases were found where one student tried to submit the flag of another student despite the significant penalties that were advertised to the students beforehand. All connected individuals got penalty points causing a lower grade for the course. The fact that students tried to cheat even fully knowing that they will very likely be discovered and punished shows how important it is to construct assignments that are personalised.

At the end of the course students were given a questionnaire to assess how useful, interesting, and difficult the lab was. The Questionnaire consisted of three 6-point Likert scale (from 0 to 5) questions. Also, students were given an open-ended question to write their comments about the labs. Out of 94 participants, 85 students answered to all the evaluation questions. The lab was rated as “quite difficult” (average rating 4.22 and 4.33 respectively for group 1 and group 2, “rather interesting” (3.33 and 3.65) and “rather useful” (3.11 and 3.26). Students in the second group found the OSINT lab to be of a similar difficulty to the first group, but slightly more interesting and useful than the first group. The first group did not benefit from the experience and shared knowledge that was available for the second group. It is considered possible that further guidance and more thorough preparation for the labs would have increased the positive feedback.

Comparing the results of each week’s performance revealed that students actively communicate with each other. This was further confirmed by anecdotal evidence from students’ feedback. Therefore, it demonstrates the importance of implementing anti-cheating mechanisms and having individual flags.

6. Future work

The difference in the results from students doing the exercise in separate weeks raises the question of whether the students in week 2 truly learned more about OSINT techniques. They may have used the information from the other students directly to access the flags without further reflection. Alternatively, they might have had less homework in other courses and therefore more time to focus on the OSINT exercise. Future research could find interesting insights by looking deeper into group dynamics regarding information sharing about homework assignments.

Students were found to have different approaches in solving the tasks. Some students tried to guess or brute force the answers while others tried to follow the step by step instruction. Both approaches were considered valid and we think that it is important to utilise a variety of means to find out the answers. A more detailed logging could provide further insights into the method that was used to capture a specific flag. Currently, it is only possible to distinguish different paths to the correct flag by analysing the flag submission log. This provided evidence that some students tried (unsuccessfully) to brute force the answers. In future, finding common patterns in students’ approaches can help predict the possible wrong approaches and help create tailored hints for specific behaviours to guide students to the correct answer. Comparing the logged solution approaches can also help evaluate the average time needed to complete the lab based on target group. Additionally, the more detailed logging can help estimating the difficulty of the flags and provide useful input for scoring different flags accordingly.

Finally, being able to sufficiently log the students’ behaviour can indicate different underlying skills. These measured skills could be mapped to a wider competency framework as described by Mäses et al (2018). More difficult tasks could indicate a higher-level skill and – depending on the exercise context – give participants more points. However, playing against the game by brute forcing the answers should be discouraged with more control mechanisms set in place in the future.

7. Conclusions

This paper introduced a novel method for creating personalised virtual labs that teach OSINT-related skills. Using self-generated certificate authority, reverse proxy, Domain Name System (DNS) cache poisoning, and redirecting Internet traffic connected to the lab machine enabled individual flags to be created for each participant. This method was implemented in a bachelor level course and proved to be an effective training platform. In addition to the technical implementation techniques, this research also investigated content design. Student behaviour and feedback was noticeably influenced by the timing of the exercise. This revealed that more guidance might be required for a beginner level course and the importance of having personalised tasks. This method for creating individual tasks for teaching OSINT-related skills has the potential to stimulate further research in this field to support the education of future cybersecurity specialists.

8. References

Bazzell, M., 2016. *Open source intelligence techniques: resources for searching and analyzing online information*. CreateSpace Independent Publishing Platform.

Best, C., 2008. Open source intelligence. *F. Fogelman-Soulié, Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining, and Their Applications to Security*, pp.331-343.

Caliskan, E., Topgul, M.O. and Ottis, R., 2017. Cyber Security Exercises: A Comparison of Participant Evaluation Metrics and Scoring Systems. *Strategic Cyber Defense: A Multidisciplinary Perspective*, *48*, p.180.

Chou, Y., 2015. *Actionable gamification: beyond points, badges, and leaderboards*. Octalysis Group, United States.

Cui, M., Cao, Z. and Xiong, G., 2018, June. How Is the Forged Certificates in the Wild: Practice on Large-Scale SSL Usage Measurement and Analysis. In *International Conference on Computational Science* (pp. 654-667). Springer, Cham.

Durumeric, Z., Kasten, J., Bailey, M. and Halderman, J.A., 2013, October. Analysis of the HTTPS certificate ecosystem. In Proceedings of the 2013 conference on Internet measurement conference (pp. 291-304). ACM.

T. J. Ellis and Y. Levy. Towards a guide for novice researchers on research methodology: Review and proposed methods. *Issues in Informing Science & Information Technology*, 6, 2009.

Ernits, M. and Kikkas, K., 2016, July. A live virtual simulator for teaching cybersecurity to information technology students. In International Conference on Learning and Collaboration Technologies (pp. 474-486). Springer, Cham.

Figueroa, S., Carías, J.F., Añorga, J., Arrizabalaga, S. and Hernantes, J., 2018, June. A RFID-based IoT Cybersecurity Lab in Telecommunications Engineering. In *2018 XIII Technologies Applied to Electronics Teaching Conference (TAEE)* (pp. 1-8). IEEE.

Fortinet Inc (2017). Threat Landscape Report Q4 2017. [online] Fortinet Inc. Available at: https://www.fortinet.com/content/dam/fortinet/assets/threat-reports/threat-report-q4-2017.pdf [Accessed 11 Oct. 2019].

Gibson, S.D., 2014. Exploring the role and value of open source intelligence. In *Open Source Intelligence in the Twenty-First Century* (pp. 9-23). Palgrave Macmillan, London.

Hamari, J., Koivisto, J. and Sarsa, H., 2014, January. Does Gamification Work? - A Literature Review of Empirical Studies on Gamification. In *HICSS* (Vol. 14, No. 2014, pp. 3025-3034).

Kaur, J. and Elsner, T., 2015. Behind the Scenes: Accepting Untrusted Certificates in a Web Browser.

Konak, A., 2018. Experiential Learning Builds Cybersecurity Self-Efficacy in K-12 Students. *Journal of Cybersecurity Education, Research and Practice*, *2018*(1), p.6.

Mäses, S., Randmann, L., Maennel, O. and Lorenz, B., 2018, July. Stenmap: framework for evaluating cybersecurity-related skills based on computer simulations. In International Conference on Learning and Collaboration Technologies (pp. 492-504). Springer, Cham.

Morschheuser, B., Hamari, J., Werder, K. and Abe, J., 2017. How to gamify? A method for designing gamification. In *Proceedings of the 50th Hawaii International Conference on System Sciences 2017*. University of Hawai'i at Manoa.

Quick, D. and Choo, K.K.R., 2018. Digital forensic intelligence: Data subsets and Open Source Intelligence (DFINT+ OSINT): A timely and cohesive mix. *Future Generation Computer Systems*, *78*, pp.558-567.

Rege, A., 2015. Multidisciplinary experiential learning for holistic cybersecurity education, research and evaluation. In *2015 {USENIX} Summit on Gaming, Games, and Gamification in Security Education (3GSE 15)*.

Schaurer, F. and Störger, J., 2013. The evolution of open source intelligence (OSINT). *Comput Hum Behav*, *19*, pp.53-56.

Son, S. and Shmatikov, V., 2010, September. The hitchhiker's guide to DNS cache poisoning. In *International Conference on Security and Privacy in Communication Systems* (pp. 466-483). Springer, Berlin, Heidelberg.

Stricek, A., 2002. A reverse proxy is a proxy by any other name. *SANS Institute InfoSec Reading Room*, *13*.

Wells, D. and Gibson, H., 2017. OSINT from a UK perspective: Considerations from the law enforcement and military domains. *Proceedings Estonian Academy of Security Sciences, 16: From Research to Security Union*, *16*, pp.84-113.

Yoshida, N., 2008. Dynamic CDN against flash crowds. In *Content Delivery Networks* (pp. 275-296). Springer, Berlin, Heidelberg.

Yuan, X., Williams, I., Kim, T.H., Xu, J., Yu, H. and Kim, J.H., 2017. Evaluating hands-on labs for teaching SQL injection: a comparative study. *Journal of Computing Sciences in Colleges*, *32*(4), pp.33-39.

9. Appendix 1 – Example Reverse Proxy Apache Configuration

The file below shows the configuration for reverse proxying port 443 traffic of the domain example.com with Apache. This configuration allows the Reverse Proxy to use the fake SSL certificates signed by the lab's CA and pass the traffic to the real website. Apache substitutes the traffic based on its conditions and replaces the specified HTML response with the unique flag generated.

```
<VirtualHost *:443>
       ServerName example.com
       SSLProxyEngine on
       SSLCertificateFile example.com.crt
       SSLCertificateKeyFile example.com.key
       SSLCACertificateFile OSINT.crt
       SSLOptions +ExportCertData
       ProxyRequests Off
       <Proxy *>
              Order allow,deny
              Allow from All
       <Proxy>
       RequestHeader unset Accept-Encoding
       "s|<div>Please Replace me |Nice Job! E{12345678901234} |i"
       FilterDeclare NEWPATHS
       FilterProvider NEWPATHS SUBSTITUTE "%{Content_Type} =~ m|^text/html|"
       FilterChain NEWPATHS
       ProxyPass / https://example.com/
       ProxyPassReverse / https://example.com/
       ErrorLog /var/log/apache2/example-ssl.local-error.log
       CustomLog /var/log/apache2/example-ssl.local-access.log combined
</VirtualHost>
```

10. Appendix 2 – Example Script for Flag Generation

The script below shows how the flag generation process works in detail for the domain example.com. Some parts of the script have been redacted to limit the possibility of brute forcing the flags. New flags are generated every time a user starts the lab, due to randomization and a low number of users, each user gets their own set of flags.

```
#!/bin/bash
domain="example.com"
#Flag number - Some domains might have multiple flags within them
flagNumber="1"
#Generate a random string only for this lab instance
#This script should run every time a user starts the lab
PHP=`which php`
$PHP generateSecret.php
#Get the randomly generated secret to create the hash
flag_secret=$(cat secret.txt)
#Generate the flag name based on domain and flag number
flag_name="flag_${domain}_task_${flagNumber}"
#Get the username of the lab participant
username=$(cat username.txt)
#Create the raw flag text that needs to be hashed
flag_raw="$username$flag_secret$flag_name"
#Create final flag from the username, flag number, domain, and the secret hash
#First 16 character of the generated hash will become the final flag
flag=$(echo -n $flag_raw | hashGenerator | awk '{ print $1 }' | head -c 16)
#put the generated flag in the site reverse proxy configuration file
#to be inserted in the user traffic
sed -i "s/$flag_name/$flag/g" ${domain}.conf
```

11. Appendix 3 – Example Script for Redirecting Domain Traffic

The code below shows how the content of a domain example.com is modified with 1.1.1.1 as an IP address. The script generates a fake certificate and redirects the users with DNS Spoofing or iptables depending on the need. This script allows a flag to be embedded without destabilizing the browsing experience of the training lab users. This bash script also relies on custom configuration and preconfigured tools, but the script below shows in general how the game design for inserting flags works.

```
#!/bin/bash
domain="example.com"
domainIP="1.1.1.1"
#Generate a CA, only issue it once
openssl req -new -config customopenssl.cnf -newkey rsa:4096 -sha256 \
       -keyout OSINT.key -x509 -days 7300 -text -out OSINT.crt
#Create the SSL key
openssl req -config openssl.cnf -new -newkey rsa:4096 -sha256 \
       -keyout "${domain}.key" -subj "/CN=${domain}" \
       -text -out "${domain}.csr"
#Sign the CSR
openssl ca -config customopenssl.cnf -in "${server}.csr" –out \
       "${server}.crt" –batch \
<(printf "\n[SAN]\nsubjectAltName=DNS:${domain},DNS:*.${domain}"))
#DNS Spoofing - Manual DNS records to redirect the users to our reverse proxy
#Only if users do not need to work with the real IP address of the domain:
echo -e 'port=53\ndomain-needed\nbogus-priv\nstrict-order \
       \nlisten-address=192.168.8.254\nno-hosts \
       \naddn-hosts=/etc/dnsmasq.hosts' >> /etc/dnsmasq.conf
echo -e '192.168.8.254 ${domain}"' >> /etc/dnsmasq.hosts
#Rediect port packets to 80 and 443 to our revese proxuy
#Only If users need to know to see the real IP address:
iptables -t nat -A PREROUTING -p tcp -d ${domainIP} --dport 80 –j \
       DNAT --to-destination 192.168.8.254:80
iptables -t nat -A PREROUTING -p tcp -d ${domainIP} --dport 443 –j \
       DNAT --to-destination 192.168.8.254:443
```

Appendix 6

VI

S. Mäses, H. Aitsam, and L. Randmann. A method for adding cyberethical behaviour measurements to computer science homework assignments. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, page 18. ACM, 2019

A method for adding cyberethical behaviour measurements to computer science homework assignments

Sten Mäses
Tallinn University of Technology
Tallinn, Estonia
sten.mases@taltech.ee

Heleri Aitsam
Tallinn University of Technology
Tallinn, Estonia
heleri.aitsam@taltech.ee

Liina Randmann
Tallinn University of Technology
Tallinn, Estonia
liina.randmann@taltech.ee

ABSTRACT

Modern computer science problems increasingly require not only technical solutions but solving complex ethical problems. In this paper, we suggest a method for adding ethical aspects more into computer science study programmes and show with a case study how it is relatively easy to add an additional ethical dimension to various technical tasks.

CCS CONCEPTS

• **Applied computing** → **E-learning**; • **Social and professional topics** → *Codes of ethics*.

KEYWORDS

cyberethics, ethical behaviour, computer science education

ACM Reference Format:
Sten Mäses, Heleri Aitsam, and Liina Randmann. 2019. A method for adding cyberethical behaviour measurements to computer science homework assignments. In *19th Koli Calling International Conference on Computing Education Research (Koli Calling '19), November 21–24, 2019, Koli, Finland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3364510.3364529

1 INTRODUCTION

Despite the vast array of complex ethical dilemmas in the field of modern computer science, there seems to be surprisingly little focus on cyberethics in educational programmes. While there are some initiatives for creating interactive ethics courses [4] and introducing ethical dilemmas into social science courses [24], the general approach tends to stay rather generic and theoretical.

At the same time, people employed in the field of computer science are increasingly faced with various ethical dilemmas. In some cases, humans are required to define their values for the machine such as programming self-driving cars to “make difficult ethical decisions in cases that involve unavoidable harm” [5]. In other cases, people might deviate from expected ethical behaviour due to carelessness [16] or optimistic bias [22]. However, those small deviances in behaviour might cause significant shortcomings in work quality and cost the whole organisation large amount of time and money to fix the situation.

We argue, that different ethical dilemmas should be introduced more widely into the computer science curricula than just being part of one ethics lecture or course. Furthermore, the students should have the opportunity to practice ethical decision making in realistic contexts that are relevant to their field.

In this paper, a simple method for adding the ethical aspects to regular computer science homework assignments is presented. This method enables to go further from just asking about the reported behaviour in theoretical contexts, and actually observe the behaviour in a simulated environment. Although the illustrating case study is conducted in the cybersecurity context, the method can easily be adjusted to other computer science courses.

2 BACKGROUND AND RELATED WORK

Ethics is an ancient discipline [14] that arguably originates already from the pre-human era [29]. The concept of computer ethics (used synonymously with cyberethics in this paper) can be traced back to mid-20th century to both fiction (e.g. [1]) and non-fiction [28]. Computer ethics was already discussed in detail by Moor in 1985 [18], but it has become increasingly relevant with machines gaining significant influence over human well-being [5].

In the effort of making computer systems safe and secure, it is not enough to establish efficient technical practices. Humans often find unethical ways how to deviate from the expected behaviour and perform in a way that requires less effort. For example, organisations might perform phishing tests on their employees for noble causes (increasing their security posture, specifying personnel training needs etc.), but without proper permissions, this activity could be considered unethical and even illegal [15]. Buchanan et al. [7] demonstrate how novel topics in the computer science field stir confusion amongst research ethics boards. They conclude that “strong ethical guidance is needed" and that it is important to “evaluate anew what ethical issues are being covered in curricula” [7].

While there is much discussion on which topics should be taught in ethics courses, there is a significant lack of research about measuring ethical behaviour in classroom context. There are experiments conducted in laboratory settings about ethical decisions [3], but the grading process in classroom setting regarding ethics is generally based on essays or questionnaires, not actual behaviour [19].

One of the few ethical aspects that is actively discussed and monitored in educational context, is plagiarism and cheating. There is an ongoing battle between those who try to cheat and those who try to understand [2] and detect cheating [20]. Despite the topics of plagiarism and cheating being important, we argue that the ethical behaviour in the field of computer science is a much wider issue.

There are many situations in cyberethics when it is difficult to define what would be an ethical way forward or whether someone

Koli Calling '19, November 21–24, 2019, Koli, Finland

ACM ISBN 978-1-4503-7715-7/19/11…$15.00
https://doi.org/10.1145/3364510.3364529

is acting in an unethical manner. For example, is writing profanities into the software source code [13] acceptable or not? Is it ethical for a software developer to balance the trade-off between the number of false positive and false negative results or should the end-users be responsible for defining all the parameters in a program they use [12]? Those questions do not have straightforward answers.

To cope with the myriad of ethical issues, it is essential that students are introduced to those kinds of questions repeatedly throughout their educational journey. Welsh et al. [27] show how small ethical transgressions are likely to lead into gradually increasing indiscretions. Therefore, it is crucial to let students play out different scenarios in a safe academic environment instead of doing it later in their professional career.

3 METHODOLOGY

The goal of this study is to demonstrate how measurements of ethical behaviour can be easily added to the existing educational tasks. Although in our study, we focus on cyberethical behaviour in cybersecurity context, it is easy to adapt this method to many computer science related courses.

To test out the suggested method for ethical behaviour measurements, the case study method was chosen. Three different homework assignments were modified to include ethical decision-making tasks. Crossover design was used in homework distribution process to minimize the influence of content of the homework to the results of the ethical behaviour (as shown in Figure 1). Crossover design also enabled to give every student the chance to do all the homework assignments without the need of an additional control group needing to skip the lecture or some homework assignments [8].

3.1 Participants

Empirical analysis was carried out as a part of a bachelor (undergraduate) level introductory cybersecurity course. 112 course participants were asked for consent to use their results in the study. 26 students did not give their consent and their data was discarded from the analysis. 86 students gave their consent and the following analysis is based on their data. There were 72 male (84%) and 14 female participants (16%). The 86 participants were randomly allocated into two groups: group X (45 students) and group Y (41 students).

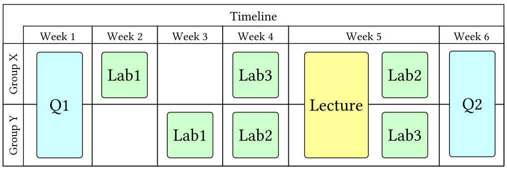
3.2 Context

The general course design was as follows: weekly lecture was followed by individual practical homework to be completed before the next lecture. Each homework contributed points to the final score (grade of the course), but it was not compulsory to complete all the homework assignments.

The following activities relevant to the ethical measurements took place over 6 weeks during spring semester of 2019:

- 1 lecture on the topic of ethics and cyberethics;
- 3 different homework assignments on the topic of cybersecurity as homework with added ethical aspects;
- survey before and after the other activities.

The general course flow is illustrated in Figure 1 showing different timeline for homework assignments according to the group. It should be noted that there were additional homework assignments (e.g., week 3 home assignment for Group X) and additional weekly lectures not related to the ethical measurements. For improved clarity, those additional activities are not discussed nor displayed in Figure 1.



	Timeline					
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Group X	Q1	Lab1		Lab3	Lab2	Q2
					Lecture	
Group Y			Lab1	Lab2	Lab3	

Figure 1: General timeline of the experiment

Q1 and Q2 in Figure 1 stand accordingly for pre- and post-questionnaire discussed further in section 3.5. Lecture stands for the lecture on ethics. The further details about the contents of the lectures are given in section 3.4. Lab1, Lab2, and Lab3 signify different homework tasks (remotely accessible virtual labs) where the ethical decision making component was added.

Virtual labs were given as individual home assignments. Participants had 5 days to remotely connect to the lab environment and complete the tasks. The topics of the technical hands-on labs were the following:

- Lab 1 - Reverse engineering;
- Lab 2 - SQL injection;
- Lab 3 - Open Source Intelligence (OSINT).

3.3 Adding ethical aspects to homework

There are many ethical aspects that can be added to a technical homework assignment. Adding an ethical aspect should usually introduce a dilemma to the student where one behaviour would be more convenient (or otherwise beneficial) and the other way would be more ethically acceptable. The dilemma is usually part of a fictional storyline connected to the task (e.g. finding an emergency fix to a problem until the responsible specialist would return from a holiday).

Schumann et al. [24] list some key elements for an ethical dilemma including the forced choice between an ethical behaviour and a more profitable (or convenient) one.

In computer science, some topics for ethical dilemmas can be the following:

- using pirated/unlicensed software;
- setting up systems with insufficient privacy controls, and security vulnerabilities;
- (not) asking permission for conducting security testing or data processing;
- fixing vulnerabilities (finding the root cause or just dealing with superficial patching);
- dealing with unethical tasks (e.g., adding malicious or illegal functionality to a system).

In our case study, we used the dilemma of asking permission. Especially in the field of cybersecurity, having a proper written permission to conduct security testing is what often differentiates legitimate specialists from illegal criminals. As part of the homework storyline, students were informed that it would be polite to ask their (fictional) boss for written permission before starting their technical tasks. If the student chose not to ask permission, then they risked losing points. The probability of losing points depended on the lab. In Lab 1 the chance of losing points, when not asking permission, was .00001, and in Lab 2 and Lab 3 .001. The aim of this was to see, if the chance of losing points was higher, whether students are more likely to ask permission or not. Students who asked permission did not receive any additional points. Therefore, the perceived risk of consequences from unethical behaviour was designed to be very low.

The permission request was to be sent to a given e-mail address and the reply was to be expected in no more than 72 hours. What the students did not know, was that actually the reply was given in 3..24 hours and that (randomly chosen) half of the replies required further action—e.g., solving a CAPTCHA. Further action had the sole purpose to make the ethical behaviour of half of the respondents even more time-consuming and therefore inconvenient. The task of asking permission was designed as an additional inconvenience that would potentially significantly limit the time available for doing the homework assignment.

Technical tasks were structured in a capture the flag format [10]. For the categorisation of student behaviour, two timestamps were collected: the time of permission given (reply to permission request), and the time of submitting the first flag (indication that the student has started the technical tasks).

3.4 Lecture about ethics

After completing the first two homework assignments with added ethical aspects, the students received one lecture on the topic of ethics (as show in Figure 1). The lecture covered the following topics:

- basics of game theory (zero-sum game, prisoner's dilemma);
- trolley problem/dilemma, comparison to ethical dilemmas of an autonomous car [5];
- defining ethics, deontological and utilitarian approaches;
- professional ethics (ethics at workplace);
- intellectual property (copyright and plagiarism).

The length of the lecture was 1.5 hours and it included several discussions of ethical case studies. For example, one fictional case study presented the case of a biotechnology company deciding to postpone privacy-related security measures to accelerate the product launch.

The students were asked to evaluate how interesting and how useful the lecture was on a 6-point Likert scale (0..5). Based on the feedback, the lecture about ethics was perceived to be very interesting (average score 4.2) and rather useful (average score 3.9).

After the lecture, the students were given their third homework including the ethical aspects.

3.5 Questionnaire design

A questionnaire was developed to measure attitude and ethical views of students before and after homework assignments containing ethical aspects and the lecture about ethics (as seen in Figure 1). The same questionnaire was used twice: before and after the other activities. The questionnaire was added to the experiment to compare the ethical behaviour during homework with the values that were self-reported in the questionnaire.

The questionnaire was compiled using selected parts from the Oxford Utilitarianism Scale (OUS) [11], the Cybersecurity Attitude Scale (CAS) [9], and Human Aspects of Information Security Questionnaire (HAIS-Q) [21]. 6-point Likert scale was used to avoid the neutral choice in the middle of the scale [6].

Firstly, the Oxford Utilitarianism Scale (OUS) [11] questions were used to measure deontological and utilitarianism views of participants. Deontology and utilitarianism can be considered to be the two main ethical theories that are compared the most [26]. The OUS does not strictly categorize people into deontologist or utilitarian; it rather gives a matter of degree [11].

The OUS consists of 9 questions, from what first five measure impartial beneficence (IB) and last four instrumental harm (IH) [11]. People who score higher on the OUS are considered with utilitarianism views and people who score lower with deontological views [11]. The subscale of instrumental harm shows the negative side of utilitarianism, how willing is a person to harm someone for greater good [11]. Impartial beneficence measures more of the positive side, the concern for others, greater good and future generations [11]. This dimension shows how much is cared for the "well-being of all sentient beings on the planet" [11]. The OUS was chosen because it covers both positive and negative side of utilitarianism.

Secondly, the Cybersecurity Attitude Scale (CAS) by Howard [9] was used to measure cyber policy adherence attitudes and perceived vulnerability to a cyberattack. The CAS consists of 10 items—first five measure policy adherence (PA) and the last five perceived vulnerability (PV). According to [9] CAS gives a better understanding, why people behave the way they do.

Policy adherence subscale shows how an individual feels about following rules and policies. Perceived vulnerability, on the other hand, measures what kind of attitude a person has towards vulnerabilities and how does (s)he perceives it [9]. As both following policies and being aware of threats and vulnerabilities are important parts of cyberethics, this scale was decided to be used in this research.

Thirdly, seven questions from The Human Aspects of Information Security (HAIS-Q) [21] were decided to use to add a third dimension – own responsibility - to the attitude scale. HAIS-Q is a questionnaire developed by Parsons et al. to measure information security awareness (ISA) [21].

HAIS-Q consists of 63 items and has seven focus areas: "Password management, Email use, Internet use, Social media use, Mobile devices, Information handling and Incident reporting" [21]. Each of these areas is additionally divided into three sub-areas – knowledge, attitude and behaviour. The items were selected from the behaviour area so that each of the seven sub-areas would be covered – one item from each of the sub-areas. These seven questions were chosen to see how students feel about taking responsibility for behaving in a secure way.

As CAS and HAIS-Q are aimed at organization employees and the items are developed from the perspective of the organization, an according storyline was created and given to students for setting the questions into a proper context. The storyline consisted of a short paragraph of descriptive text, including a part that asked the student to imagine working in a small company with 15 other people. Afterwards a simple attention-checking question was asked to improve that trustworthiness of the questionnaire answers—how many people were working in this fictional company?

4 RESULTS

86 students were divided into two groups. All of them completed the pre-questionnaire (Q1). 8 students failed in the attention check (a simple question that is easy to answer wrong if the question text is not read properly). These students' results were also removed from the pre-questionnaire analysis. Also, students, who did not participate in the lecture were removed from the analysis. In the end, 65 students' pre-questionnaire results were analysed.

21 students (32%) were with utilitarianism views and 44 (68%) deontological views, based on the OUS. 25% of the students in group X, and 41% in group Y were classified to be with utilitarianism views. Over half of the students (63%, 41 students) scored low on the impartial beneficence scale and also on the instrumental harm scale (65%, 42 students). All the students felt that following policies and rules is important.

Perceived vulnerability answers show somewhat different results. 9 students (14%) did not discern their vulnerability very highly. On HAIS-Q, 1 (2%) student did not value own responsibility highly.

In the end of the study, the post-questionnaire (Q2) was conducted that was identical to Q1. 81 students gave responses. 8 students failed the attention check question. These students' results were removed from the post-questionnaire analysis to improve the validity of the experiment. Results of the students, who did not participate in the lecture were removed. In the end, 66 students' post-questionnaire results were analysed. 21 students (32%) were with utilitarianism views and 45 (68%) deontological views, based on the OUS. Over half of the students (65%, 43 students) scored low on the impartial beneficence scale and also low on the instrumental harm scale (64%, 42 students).

To measure students' activity in the virtual hands-on labs, they were classified, based on their actions, into six groups (G1..G6):

G1 Students, who asked permission and did not start the lab before receiving one.
G2 Students, who asked permission but started the lab before they received it.
G3 Students, who asked permission, but after receiving answer requiring further action, they did not reply—hence they did not receive permission to start the lab but nevertheless did so.
G4 Students, who asked permission and received permission, but did not start the lab (did not get results).
G5 Students, who did not ask permission but started the lab.
G6 Students, who did not ask permission nor started the lab.

Student distribution based on their behaviour is displayed in Table 1.

Spearman's rank-order correlation was calculated between questionnaires and behaviour in the labs. No significant correlation was observed.

Table 1: Student behaviour per lab

	G1	G2	G3	G4	G5	G6
X-Lab1	9	15	1	2	5	7
Y-Lab1	0	13	5	0	14	2
X-Lab3	1	5	2	20	1	10
Y-Lab2	2	16	4	2	9	1
X-Lab2	6	16	4	3	7	3
Y-Lab3	10	11	4	0	8	1

Table 2: Cronbach's alphas for pre- and post-questionnaire categories

	IB	IH	PA	PV	HAIS-Q
Q1	0.68	0.75	0.77	0.86	0.48
Q2	0.72	0.74	0.89	0.90	0.75

5 DISCUSSION

Table 2 lists Cronbach's alphas per questionnaire category for both pre-questionnaire (Q1) and post-questionnaire (Q2). Cohen et al. [8] consider alpha values under 0.60 unreliable. The results indicate that in general the questionnaire was a good fit to measure ethical views and cybersecurity attitude. When comparing to other subscales the improvement in HAIS-Q items subscale show unexpectedly big change from 0.48 in pre-questionnaire to 0.75 in the post-questionnaire. This indicates a likely change in the mindset of students. It might be caused by the increase in the understanding of questions or students thinking more carefully, when answering.

Interestingly, no significant Spearman's rank-order correlation was observed between questionnaires and behaviour in the labs. That could indicate that one of the measurements (either the questionnaire or behaviour in the simulated ethical dilemma) is invalid or that the answers to the questionnaire are not strongly connected to the actual behaviour. Criticism towards behavioural models such as Theory of Planned Behaviour [25] seems to support the possibility that the reported behaviour (or attitude and other parameters) based on questionnaire results might not be strongly connected to the actual behaviour. Nevertheless, it could be argued that the student behaviour in the gamified classroom context is also not completely representative of the actual behaviour (e.g., due to Hawthorne effect [17]). While work sample testing has proven to be valid predictor of job performance [23], the interpretation of ethical behaviour measurements in gamified simulations still needs further research.

For group Y, it is visible that the results improved with each lab. When in the first lab only 38% of students got permission, then in the final lab, the number had increased to 62%. Also, the number of students, who waited before starting the lab for permission (G1) increased. In the first lab (Lab1) not a single student waited before receiving permission (G2), in the second lab (Lab2) 6% waited for permission. In the third lab (Lab3), after the cyberethics lecture was held, 29% of the students waited before starting the lab. This gives some indication that cyberethics lecture had a positive effect on the cyberethical behaviour of students.

Contrastingly, the results for group X were different. The number of students, who got permission, in the first lab (Lab1) was 67%, in the second lab (L3) this percentage stayed the same and then decreased to 64% in the third lab (L2). The reason for this might be, that in group X were more students, who already knew at the beginning of the experiment about cyberethics behaviour, but this is not very likely, because the difference between first lab results is big, but in pre-questionnaire, the group difference is not so evident. The gap between the groups X and Y comes clearly out from the results of the first lab. When in group X 67% of the students got permission in the first lab then in group Y 38% of the students got permission. Therefore, the reason for this might be the different timing. Group X did the first lab right after the pre-questionnaire, compared to group Y, who did the first lab a week later, and it might have affected the result of the lab (especially considering potential unsupervised and uncontrollable communication between the groups).

5.1 Limitations and future work

This case study was conducted to demonstrate a possible method for adding ethical dimension to technical computer science homework tasks in classroom context. While the method gives interesting additional information about the cyberethical behaviour, further research is needed for comparing the effectiveness and validity of different approaches. As shown, parameters such as timing can have significant impact on measured behaviour.

Furthermore, the labs were conducted in a gamified context and it could be argued, that in actual situations the behaviour would be different. Therefore, the results should be interpreted with caution.

6 CONCLUSIONS

In educational contexts, ethical issues have become increasingly important. The present study has demonstrated a how it is possible to add an ethical behaviour assessment to computer science homework. Measuring ethical behaviour enables to improve the related teachings and facilitates interesting future research on that topic. Although the results of such measurements are very context dependent and should be interpreted with caution, they provide good basis for further discussions both in classrooms and among the scientific community.

REFERENCES

[1] Isaac Asimov. 1950. *I, Robot.* Gnome Press.

[2] Claudio Barbaranelli, Maria L Farnese, Carlo Tramontano, Roberta Fida, Valerio Ghezzi, Marinella Paciello, and Philip Long. 2018. Machiavellian ways to academic cheating: A mediational and interactional model. *Frontiers in psychology* 9 (2018).

[3] Rachel Barkan, Shahar Ayal, and Dan Ariely. 2015. Ethical dissonance, justifications, and moral behavior. *Current Opinion in Psychology* 6, DEC (2015), 157–161.

[4] Jane Blanken-Webb, Imani Palmer, Sarah-Elizabeth Deshaies, Nicholas C Burbules, Roy H Campbell, and Masooda Bashir. 2018. A Case Study-based Cybersecurity Ethics Curriculum. In *2018 USENIX Workshop on Advances in Security Education (ASE 18).*

[5] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.

[6] James Dean Brown. 2000. What issues affect Likert-scale questionnaire formats. *Shiken: JALT Testing & Evaluation SIG Newsletter* 4, 1 (2000).

[7] Elizabeth Buchanan, John Aycock, Scott Dexter, David Dittrich, and Erin Hvizdak. 2011. Computer science security research and human subjects: Emerging considerations for research ethics boards. *Journal of Empirical Research on Human Research Ethics* 6, 2 (2011), 71–83.

[8] Louis Cohen, Lawrence Manion, and Keith Morrison. 2002. *Research methods in education.* routledge.

[9] David J Howard. 2018. Development of the Cybersecurity Attitudes Scale and Modeling Cybersecurity Behavior and its Antecedents. (2018). Master thesis in University of South Florida.

[10] Haomiao Huang, Jerry Ding, Wei Zhang, and Claire J Tomlin. 2014. Automation-assisted capture-the-flag: A differential game approach. *IEEE Transactions on Control Systems Technology* 23, 3 (2014), 1014–1028.

[11] Guy Kahane, Jim AC Everett, Brian D Earp, Lucius Caviola, Nadira S Faber, Molly J Crockett, and Julian Savulescu. 2017. Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. (2017).

[12] Felicitas Kraemer, Kees Van Overveld, and Martin Peterson. 2011. Is there an ethics of algorithms? *Ethics and Information Technology* 13, 3 (2011), 251–260.

[13] Juho Leinonen and Arto Hellas. 2017. Thought crimes and profanities whilst programming. In *Proceedings of the 17th Koli Calling International Conference on Computing Education Research.* ACM, 148–152.

[14] Alasdair MacIntyre. 1998. *A short history of ethics: a history of moral philosophy from the Homeric Age to the twentieth century.* University of Notre Dame Press, Notre Dame, Ind.

[15] Sten Mäses, Kristjan Kikerpill, Kaspar Jüristo, and Olaf Maennel. 2019. Mixed Methods Research Approach and Experimental Procedure for Measuring Human Factors in Cybersecurity Using Phishing Simulations. In *18th European Conference on Research Methodology for Business and Management Studies.* 218.

[16] Thulani Mashiane and Elmarie Kritzinger. 2018. Cybersecurity Behaviour: A Conceptual Taxonomy. In *IFIP International Conference on Information Security Theory and Practice.* Springer, 147–156.

[17] Jim McCambridge, John Witton, and Diana R Elbourne. 2014. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology* 67, 3 (2014), 267–277.

[18] James H Moor. 1985. What is computer ethics? *Metaphilosophy* 16, 4 (1985), 266–275.

[19] Barbara Moskal, Keith Miller, and LA King. 2002. Grading essays in computer ethics: rubrics considered helpful. In *ACM SIGCSE Bulletin*, Vol. 34. ACM, 101–105.

[20] Julia Opgen-Rhein, Bastian Küppers, and Ulrik Schroeder. 2018. An Application to Discover Cheating in Digital Exams. In *Proceedings of the 18th Koli Calling International Conference on Computing Education Research.* ACM, 20.

[21] Kathryn Parsons, Dragana Calic, Malcolm Pattinson, Marcus Butavicius, Agata McCormac, and Tara Zwaans. 2017. The human aspects of information security questionnaire (HAIS-Q): two further validation studies. *Computers & Security* 66 (2017), 40–51.

[22] Hyeun-Suk Rhee, Young U Ryu, and Cheong-Tag Kim. 2012. Unrealistic optimism on information security management. *Computers & Security* 31, 2 (2012), 221–232.

[23] Philip L Roth, Philip Bobko, and Lynn A McFarland. 2005. A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology* 58, 4 (2005), 1009–1037.

[24] Paul L Schumann, Timothy W Scott, and Philip H Anderson. 2006. Designing and introducing ethical dilemmas into computer-based business simulations. *Journal of Management Education* 30, 1 (2006), 195–219.

[25] Falko F. Sniehotta, Justin Presseau, and Vera Araújo-Soares. 2014. Time to retire the theory of planned behaviour. *Health Psychology Review* 8, 1 (2014), 1–7.

[26] Richard A Spinello and Herman T Tavani. 2001. The Internet, ethical values, and conceptual frameworks: an introduction to Cyberethics. *ACM SIGCAS Computers and Society* 31, 2 (2001), 5–7.

[27] David T Welsh, Lisa D Ordóñez, Deirdre G Snyder, and Michael S Christian. 2015. The slippery slope: How small ethical transgressions pave the way for larger future transgressions. *Journal of Applied Psychology* 100, 1 (2015), 114.

[28] Norbert Wiener. 1948. *Cybernetics, or Control and Communication in the Animal and the Machine.* The Technology Press.

[29] Peter Wohlleben. 2017. *The inner life of animals: love, grief, and compassion: surprising observations of a hidden world.* Greystone Books David Suzuki Institute, Vancouver Berkeley.

Appendix 7

VII

S. Mäses, O. Maennel, and S. Sütterlin. Using competency mapping for skill assessment in an introductory cybersecurity course. In *Educating Engineers for Future Industrial Revolutions - Proceedings of the 23rd International Conference on Interactive Collaborative Learning (ICL2020)*. Springer, 2020

Using competency mapping for skill assessment in an introductory cybersecurity course

Sten Mäses[1], Olaf Maennel[1], and Stefan Sütterlin[2]

[1] Tallinn University of Technology, Tallinn, Estonia,
sten.mases@taltech.ee, olaf.maennel@taltech.ee
[2] Ostfold University College, B R A Veien 4, Halden, Norway
stefan.sutterlin@hiof.no

Abstract. Various courses and trainings aim to teach cybersecurity with only few measures for learning outcomes. Existing metrics such as general grades do not reflect the the various competencies and specific outcomes that make up a course. More specific measures targeting competencies are needed: this differentiated approach must reflect course specificities, but also be general enough to be transferable to other courses.

This study presents a competency mapping approach based on the NIST NICE framework, demonstrated in a cybersecurity course on bachelor level. The approach enables to evaluate specific competencies and their relevance to various cybersecurity job roles. Additionally, this type of competency mapping can provide useful insights for designing and managing the course content in general.

Keywords: cybersecurity, skill assessment, computer science education, competency mapping, NICE framework

1 Introduction

Since the Third Industrial Revolution, which is generally attributed to computerization and web-based interconnectivity [1], there is a growing demand for skills to operate various interconnected computing machines in an efficient and secure way. However, there is a reported shortage in the job market for people with cybersecurity competencies. Higher education systems are facing increasing pressure to produce graduates with relevant practical skills and competencies. This could be achieved by focusing more on competencies that are relevant in the job market [2]. In addition to giving a general course grade that reflects the summative level of achieving the learning objectives, a more detailed competency-based feedback could be provided. This would enable earlier identification of problematic areas for each student to reduce dropout rate and increase the quality of education. Also, it would potentially help to discover the so-called hidden talents – people who have difficulties in some cybersecurity-related topic might not realise that they could be suitable for a different cybersecurity job.

The goal of this study is to systematically map the results of various learning activities to a specific competency framework, therefore creating an overview of the student's competencies. That enables a better understanding of the student's current performance for both the student and the teacher. Otherwise, students with very different competency profiles across a variety of task types and demands might get identical resulting grades that do not help them to further analyse their strengths and weaknesses [3]. Without competency mapping, it is much more difficult to understand what work role or topic area best fits the student. Established individual competency profiles, on the other hand, can contribute to early detection and intervention targeting insufficient competency areas and identify or build upon existing strength to facilitate specialisation.

This paper—aligned with the general shift towards the competency-based education [4]—discusses a case study of implementing competency mapping in an introductory cybersecurity course. The process follows design science research methodology [5] describing different design and implementation concerns. Questions of the knowledge tests and practical homework tasks are assigned difficulty levels, and mapped into the NICE framework [6]. The resulting competency profile enables deeper insights in both individual and course level.

2 Related work

As shown by Frezza et al. [7], the term competency and competence are somewhat ambiguous. The current paper is aligned with research considering the competency as a set of knowledge, skills, and dispositions [7–9]. More specifically, the focus is on measurements of knowledge and skills. The competency mapping is well aligned with the general shift towards a competency-based pedagogy [4]. Having a systematic approach for identifying and evaluating competencies—that is the basis for any competency-based course.

A considerable amount of literature has been published on curriculum mapping [10]. Ajanovski [11] introduces a visual domain exploration tool that shows the personal progress over a reference body of knowledge areas. Gluga [12] explores the idea of mapping the detailed degree requirements to the individual courses, the assessments and each student's actual performance on assessment tasks. Auvinen et al. have created STOPS (Software for Target-Oriented Personal Syllabus) [13] that enables to create a graph showing how the topics taught in various courses are connected to each other, and what each course contributes to the goals that the student has selected. Harris and Patten map various cybersecurity topics to courses in the IT curriculum [14]. For determining the expected proficiency level, they combine revised Bloom's Taxonomy [15] and Webb's Depth of Knowledge Model [16]. However, there is little effort on looking deeper than the course level learning outcomes. While Gluga [12] mentions mapping to students' actual performance, there seems to be no further published research based on real data at the moment.

Another set of research is looking at interdependencies of competencies. Frezza et al. [7] define a competency-based framework for learning to provide a structured approach to deriving and describing competencies. Deng et al. [17] use machine learning (natural language processing) to construct a cybersecurity knowledge graph from Wikipedia database dump. Current study explores the competency mapping that can add an extra dimension to the study of interdependencies, because it enables to analyse knowledge and skill proficiency evaluations based on individual results opposed to the course level analysis.

3 Methodology

This research is using design science research methodology (DSRM) process [5] to present the process of competency mapping for skill assessment. The DSRM process consists of 6 steps: problem identification and motivation, defining the objectives of a solution, design and development, demonstration, evaluation, and communication [5]. The last step—communication—is done by this paper in general. The first 5 steps are discussed in the following sections.

3.1 Problem identification and motivation

To address the shortage of cybersecurity competencies in the job market, the educational processes must become more efficient. Practical exercises are often part of a course, giving a fixed percentage of the general grade. The final grade for a course is often considered as a performance indicator for a student. However, grade inflation is happening in many countries [3]. Universities are faced with the dilemma of whether it is more important to provide maximum amount of students a required set of baseline competencies or to ensure normal distribution of the grades. Normal distribution of the grades would help better compare the proficiency levels of different students by avoiding the more concentrated grades at the upper tail [3]. At the same time, letting the performance of others to influence individual grades would further complicate the interpretation of the result as it becomes highly context-dependent.

In general, the grading of courses and the scoring of cybersecurity exercises suffer from a similar shortcoming—lack of meaningful interpretation of the result [18]. While it could be argued that a higher score in a cybersecurity competition and a higher grade in a cybersecurity course both signify a higher proficiency in the topic, a deeper analysis of underlying competencies is often difficult if not impossible. As an example, let us imagine Alice and Bob both receiving a grade 7 (out of 10) in a cybersecurity course. It is usually not possible to deduct from the grade whether Alice is more proficient regarding a specific learning outcome than Bob or not. That is because the learning outcomes of the course are rarely evaluated individually. Furthermore, there is little distinguishing between the scores from theoretical knowledge tests and practical skills evaluations. So it is not possible to see whether a student has demonstrated good results in theory, in practice, or in both.

3.2 Objectives of a solution

The goal is to achieve more systematic overview of individual competencies of each student (in the classroom education context). The method for achieving this goal should include the following:

1. Measured competencies should be structured systematically into categories and/or clusters.
2. Meaningful proficiency levels—e.g., higher and lower level tasks should have a clearly defined qualitative difference, not just arbitrary "difficulty level".
3. Differentiation between the evaluation scores of knowledge and skills.

3.3 Design and development

Following the objectives set earlier, the design process went through the following phases:

1. selecting suitable cybersecurity competency framework listing relevant knowledge and skills and/or topic areas;
2. selecting suitable educational taxonomy for meaningful proficiency levels;
3. mapping the tasks in the course to the selected competency framework, and assigning the proficiency levels according to the educational taxonomy.

Cybersecurity competency frameworks While there are multiple initiatives gathering a list of topics relevant to cybersecurity [8, 19], the most prominent ones tend to be the following three. First, a set of curricular recommendations in cybersecurity education (CSEC2017 [20], CSEC2020 [21]), released by a joint task force led by the Association for Computing Machinery and the IEEE Computer Society. Second, Centers Of Academic Excellence In Cyber Defense Education (CAE-CD) [22] created in the USA by the National Security Agency and the Department of Homeland Security. To be recognised as CAE, a university must have a designated curriculum covering a required set of knowledge units [22]. Third, NIST National Initiative for Cybersecurity Education Cybersecurity Workforce Framework (NICE framework) [6] listing knowledge, skills, and abilities that are needed for cybersecurity-related jobs.

In the current case, the NICE framework [6] was used. First, because it is widely used both in research and in private sector. Secondly, because it provides the extensive list of knowledge and skills suitably categorised into cybersecurity work roles—enabling to compare the student performance from a perspective of a potential future work role. The NICE Framework is comprised of the following components:

- 7 high level categories: Securely Provision (SP), Operate and Maintain (OM), Oversee and Govern (OV), Protect and Defend (PR), Analyze (AN), Collect and Operate (CO), Investigate (IN).
- 33 specialty areas: Risk Management (RSK), Software Development (DEV), Systems Architecture (ARC), Technology R&D (TRD), Systems Requirements Planning (SRP), Cyber Investigation (INV) etc...

- 52 work roles: Software Developer, Data Analyst, Program Manager, Vulnerability Assessment Analyst etc. . .

Each high level category comprises several specialty areas. Each specialty area contains multiple work roles. Each work role is connected to specific knowledge, skill, ability, and task items.

Educational taxonomies Several educational taxonomies aim to provide a shared language for describing learning outcomes and performance in assessments. The revised Bloom's taxonomy [23] is widely used. Fuller et al. [24] have thoroughly analysed various taxonomies in computer science context and propose a two dimensional adaptation of Bloom's taxonomy—the matrix taxonomy. Niemierko's "ABC" taxonomy [25] described by Strachanowska [26] and Fuller et al. [24] was chosen for the current study. It aligns well with the concept of distinguishing knowledge from skills and at the same time limits the complexity of proficiency levels. A more complex taxonomy requires deeper analysis of the particular context of specific learning outcomes and the classification can become problematic [24]. Niemierko's "ABC" taxonomy is generally aligned to Bloom's taxonomy, just combining the three highest Bloom categories into one. It is also somewhat similar to the single-loop and double-loop learning [27]. Single-loop signifies the repetition of the same strategy to solve a problem and double-loop emphasises critical self-reflection to generate creative solutions to problems [28].

Mapping the tasks to competencies and proficiency levels 15 knowledge tests and 12 homework assignments of the course were analysed. Knowledge tests were divided into questions and each question was mapped to relevant knowledge unit of the NICE framework. In the current study, all 589 knowledge and 365 skill items listed in the NICE framework master KSA (knowledge, skills, abilities) list were used. For mapping 111 questions used during the course to 589 knowledge items in NICE framework, an initial filtering of relevant knowledge items was conducted. Initial filtering identified 75 knowledge items relevant to the course.

Each question was assigned difficulty level 1 or 2 based on Niemierko's "ABC" taxonomy categories A and B (see Table 1). An example of level 1 question: *What does the letter A in CIA triad stand for? (Right answer: availability.)* An example of level 2 question: *Mary adds her digital signature to her essay. Which security aspect does it help to protect? (Right answer: integrity.)* Answering correctly the level 2 question results in a proficiency point in both level 1 and level 2.

Tasks from the practical homework assignments were connected to relevant skills in NICE framework. Each task was then assigned a difficulty level 1 or two based on Niemierko's "ABC" taxonomy categories C and D. Difficulty levels determining the proficiency are summarised in Table 1.

The summary of the designed process is given in Figure 1.

Table 1. Difficulty levels based on Niemierko's "ABC" taxonomy [25] categories

Difficulty	Competency component	ABC taxonomy category
1	knowledge	A. Remembering the knowledge
2	knowledge	B. Understanding the knowledge
1	skills	C. Application of knowledge in typical problem situations
2	skills	D. Application of knowledge in unfamiliar problem situations

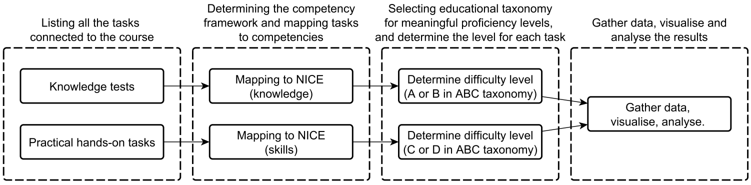

Listing all the tasks connected to the course
Determining the competency framework and mapping tasks to competencies
Selecting educational taxonomy for meaningful proficiency levels, and determine the level for each task
Gather data, visualise and analyse the results
Knowledge tests
Practical hands-on tasks
Mapping to NICE (knowledge)
Mapping to NICE (skills)
Determine difficulty level (A or B in ABC taxonomy)
Determine difficulty level (C or D in ABC taxonomy)
Gather data, visualise, analyse.

Fig. 1. Mapping process—sections with dashed lines represent the generic steps for the described competency mapping method, and sections with solid lines represent the steps taken during the current case study.

3.4 Demonstration

The case study was conducted in an introductory undergraduate level cybersecurity course with 106 participants. Performance scores were gathered from 15 knowledge tests (that took place in Kahoot[3] environment during weekly lectures) and 12 homework assignments. There were three types of homework assignments: 1 essay (evaluated by peers based on detailed criteria), 3 traditional assignments (manually graded report and/or software solution), and 8 virtual labs in a remotely accessible cloud-based environment (automatically scored).

Knowledge was assessed based on binary evaluations of Kahoot test answers (right or wrong) of each student. Each right answer gave 1 or 2 points according to its difficulty level (Niemierko's "ABC" taxonomy [25] categories A and B), contributing to the total score of related NICE knowledge units.

Skills were assessed based on homework assignments where the maximum possible scores were ranging from 4 to 10 points according to the estimated time investment and perceived importance by the teacher. To unify the skill proficiency measurements, each homework was assigned with thresholds for achieving difficulty levels 1 and 2 based on Niemierko's "ABC" taxonomy categories C and D. For example, the reverse engineering virtual lab had the threshold of at least 1 point (out of 4) for achieving level 1 proficiency and the threshold of at least 3 points (out of 4) for achieving both level 1 and level 2 proficiency.

[3] https://kahoot.com

3.5 Evaluation

In general, the created method managed to fulfil the set criteria. Measured competencies were mapped to NICE framework providing a well-structured systematic overview. Proficiency levels following the guidelines of Niemierko's "ABC" taxonomy [25] provided a good balance between simplicity and meaningful interpretation. Niemierko's "ABC" taxonomy enabled also to distinguish knowledge and skill evaluation scores.

Figure 2 compares the results of Alice and Bob based on NICE categories and specialty areas. A detailed look into performance per specific area can help finding potentially interesting and suitable areas. For example, it is visible that Bob was showing good results (skill level 2) in Cybersecurity Defense Analysis (CDA), and Alice in Exploitation Analysis (EXP). While those insights are based on initial data and should be carefully analysed further, they illustrate how competency mapping can provide relevant ideas for future skill development.

Furthermore, the mapping process itself provided useful insights regarding the course. Out of 589 knowledge items of NICE framework, only 75 were considered to be relevant to the existing course. The rest of the knowledge items provide a good basis for considering additional topics to be included in the course in the future. Also, out of the 75 relevant knowledge items, only 44 were actually measured by the quiz—demonstrating a significant gap between the topics covered by the lectures, and later knowledge measurements. For example *K0110—Knowledge of adversarial tactics, techniques, and procedures* was mentioned in several lectures, but never included in a quiz.

Additionally, the mapping provided possibility to see which NICE specialty areas were used the most during the course assessments. For example, Securely Provision (SP) category had a strong focus during the course while Investigate (IN) category was covered less (illustrated by the right side of Figure 2).

Nevertheless, it should be noted that NICE framework is a high-level framework not going into details regarding skills and knowledge items. Therefore, the mapping might not always adequately reflect the coverage of the concept. For example, let us say that a student has completed a technical task mapped to NICE skill *S0130—Skill in writing scripts using R, Python, PIG, HIVE, SQL, etc.* From seeing the skill, it is not possible to deduct what programming language was used. Furthermore, the level of proficiency is difficult to determine as the scope of "writing scripts" can be understood differently. However, some sort of ambiguity is inevitable, because defining a competency with infinite accuracy is not feasible (at least in most practical contexts).

To sum up, the mapping method fulfilled its objectives and provided multiple additional insights as well as ideas for future research.

4 Discussion and future work

Mapping all the tasks included in the course to a competency framework is a time consuming, but rewarding process. Thinking in terms of competencies and specific measurements helps to focus the course materials, and assessments.

However, many so-called soft or non-technical skills are more difficult to be measured in a classroom context, and especially in a scalable way (e.g., using an automated virtual lab). For example, NICE framework skills such as *S0355—Skill in negotiating vendor agreements and evaluating vendor privacy practices* or *S0356—Skill in communicating with all levels of management including Board members* are difficult to simulate and measure in any classroom context. Nevertheless, the importance of such transferable skills (such as critical thinking, communication, collaboration and creativity [29]) is increasingly brought out [30, 18, 31]. Although there have been attempts to measure non-technical competencies such as cyberethical behaviour [32] or teamwork skills [33, 34], more research is required for specifying relevant non-technical competencies and developing assessment methods suitable for computer science context.

It is not possible to cover all the relevant cybersecurity topics in one course. There have been initiatives for bringing in cybersecurity topics into other computer science courses [35], social science courses [36, 37] and liberal arts curriculum [38]. Aspirationally, a unified competency mapping could help track students' competencies throughout different courses. Cybersecurity competency frameworks could be combined with work on other computer science fields already having existing body of knowledge [39] and various taxonomies [40]. Similar approach could be used in other fields as well, potentially even in areas not directly connected to computer science.

This paper did not go deep into visualisation or usability aspects of competency map. Various user experience aspects of the competency map application will provide several future research opportunities. Approaches such as LATUX workflow [41] can be used for designing, validating, and deploying learning analytics visualizations Observing competency map assessments through time can give valuable insights about learning paths [42].

5 Conclusions

Competency mapping for skill assessment can provide valuable input for various learning processes. An exploratory case study of competency mapping was carried out in a cybersecurity course where different learning evaluations were connected to NICE framework. Preliminary results show that this kind of approach can be beneficial for planning and structuring a course's contents with regard to the students' reached levels of competencies throughout the course.

Although this paper focuses on cybersecurity context, a similar approach can be used in other computer science and engineering areas. In the future, similar competency mapping conducted across different courses has the potential to help curricular design and construct better learning analytics environments [9]. When applying a similar approach to several courses, then a more comprehensive competency map can be compiled. Aspirationally, a granular competency map could complement the traditional higher education diploma and provide valuable insights for both the graduate and her future employer. For now, the detailed process of competency mapping gives good basis to try similar approaches in other courses.

References

1. B. E. Penprase, “The fourth industrial revolution and higher education,” in *Higher education in the era of the fourth industrial revolution*, pp. 207–229, Springer, 2018.
2. M. Gerstein and H. H. Friedman, “Rethinking higher education: Focusing on skills and competencies,” *Gerstein, Miriam and Hershey H. Friedman (2016),“Rethinking Higher Education: Focusing on Skills and Competencies,” Psychosociological Issues in Human Resource Management*, vol. 4, no. 2, pp. 104–121, 2016.
3. M. Nordin, G. Heckley, and U. Gerdtham, “The impact of grade inflation on higher education enrolment and earnings,” *Economics of Education Review*, vol. 73, p. 101936, 2019.
4. M. Henri, M. D. Johnson, and B. Nepal, “A review of competency-based learning: Tools, assessments, and recommendations,” *Journal of engineering education*, vol. 106, no. 4, pp. 607–638, 2017.
5. K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.
6. W. Newhouse, S. Keith, B. Scribner, and G. Witte, “National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework,” *NIST Special Publication 800-181*, 2017.
7. S. Frezza, M. Daniels, A. Pears, Å. Cajander, V. Kann, A. Kapoor, R. McDermott, A.-K. Peters, M. Sabin, and C. Wallace, “Modelling competencies for computing education beyond 2020: a research based approach to defining competencies in the computing disciplines,” in *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pp. 148–174, ACM, 2018.
8. A. Parrish, J. Impagliazzo, R. K. Raj, H. Santos, M. R. Asghar, A. Jøsang, T. Pereira, and E. Stavrou, “Global perspectives on cybersecurity education for 2030: a case for a meta-discipline,” in *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pp. 36–54, ACM, 2018.
9. M. Sabin, H. Alrumaih, and J. Impagliazzo, “A competency-based approach toward curricular guidelines for information technology education,” in *2018 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1214–1221, IEEE, 2018.
10. F. Rawle, T. Bowen, B. Murck, and R. Hong, “Curriculum mapping across the disciplines: differences, approaches, and strategies,” *Collected Essays on Learning and Teaching*, vol. 10, pp. 75–88, 2017.
11. V. V. Ajanovski, “Body of knowledge explorer: Long-term student guidance across the computer-science domain,” in *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, p. 5, ACM, 2019.
12. R. Gluga, “Long term student learner modeling and curriculum mapping,” in *International Conference on Intelligent Tutoring Systems*, pp. 227–229, Springer, 2010.
13. T. Auvinen, J. Paavola, and J. Hartikainen, “Stops: a graph-based study planning and curriculum development tool,” in *Proceedings of the 14th Koli Calling International Conference on Computing Education Research*, pp. 25–34, ACM, 2014.

14. M. A. Harris *et al.*, "Using bloom's and webb's taxonomies to integrate emerging cybersecurity topics into a computic curriculum," *Journal of Information Systems Education*, vol. 26, no. 3, p. 4, 2019.
15. D. R. Krathwohl and L. W. Anderson, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Longman, 2009.
16. N. L. Webb, "Research monograph no. 6: Criteria for alignment of expectations and assessments in mathematics and science education," *Washington, DC: Council of Chief State School Officers*, 1997.
17. Y. Deng, D. Lu, D. Huang, C.-J. Chung, and F. Lin, "Knowledge graph based learning guidance for cybersecurity hands-on labs," in *Proceedings of the ACM Conference on Global Computing Education*, pp. 194–200, ACM, 2019.
18. S. Mäses, L. Randmann, O. Maennel, and B. Lorenz, "Stenmap: framework for evaluating cybersecurity-related skills based on computer simulations," in *International Conference on Learning and Collaboration Technologies*, pp. 492–504, Springer, 2018.
19. J. Hallett, R. Larson, and A. Rashid, "Mirror, mirror, on the wall: What are we teaching them all? characterising the focus of cybersecurity curricular frameworks," in *2018 {USENIX} Workshop on Advances in Security Education ({ASE} 18)*, 2018.
20. D. L. Burley, M. Bishop, S. Buck, J. J. Ekstrom, L. Futcher, D. Gibson, E. K. Hawthorne, S. Kaza, Y. Levy, H. Mattord, and A. Parrish, *Cybersecurity Curricula 2017*, vol. 1. 2017.
21. C. Tang, C. Tucker, C. Servin, M. Geissler, M. Stange, N. Jones, J. Kolasa, A. Phillips, L. Piskopos, and P. Schmelz, "Cybersecurity curricular guidance for associate-degree programs," tech. rep., Association for Computing Machinery (ACM) Committee for Computing Education in Community Colleges (CCECC), 2020.
22. NSA, "CAE requirements and resources." http://www.iad.gov/nietp/CAERequirements.cfm, 2020. Accessed: 2020-01-07.
23. C. W. Starr, B. Manaris, and R. H. Stalvey, "Bloom's taxonomy revisited: specifying assessable learning objectives in computer science," in *ACM SIGCSE Bulletin*, vol. 40, pp. 261–265, ACM, 2008.
24. U. Fuller, C. G. Johnson, T. Ahoniemi, D. Cukierman, I. Hernán-Losada, J. Jackova, E. Lahtinen, T. L. Lewis, D. M. Thompson, C. Riedesel, *et al.*, "Developing a computer science-specific learning taxonomy," *ACM SIGCSE Bulletin*, vol. 39, no. 4, pp. 152–170, 2007.
25. B. Niemierko and W. S. i Pedagogiczne, *ABC testów osiagnieć szkolnych (in Polish).* Wydawnictwa Szkolne i Pedagogiczne, 1975.
26. I. Strachanowska, "Taxonomy abc setting operational teaching objectives in the pedagogical process of training english philology students," 1997.
27. C. Argyris, "Teaching smart people how to learn," *Harvard business review*, vol. 69, no. 3, 1991.
28. K. Kiili, "Foundation for problem-based gaming," *British journal of educational technology*, vol. 38, no. 3, pp. 394–404, 2007.
29. Y. Harari, *21 lessons for the 21st century.* New York: Spiegel & Grau, 2018.
30. M. Stockman, "Infusing social science into cybersecurity education," in *Proceedings of the 14th annual ACM SIGITE conference on Information technology education*, pp. 121–124, ACM, 2013.

31. K. S. Jones, A. S. Namin, and M. E. Armstrong, “The core cyber-defense knowledge, skills, and abilities that cybersecurity students should learn in school: Results from interviews with cybersecurity professionals,” *ACM Transactions on Computing Education (TOCE)*, vol. 18, no. 3, p. 11, 2018.
32. S. Mäses, H. Aitsam, and L. Randmann, “A method for adding cyberethical behaviour measurements to computer science homework assignments,” in *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, pp. 1–5, 2019.
33. E. Britton, N. Simper, A. Leger, and J. Stephenson, “Assessing teamwork in undergraduate education: a measurement tool to evaluate individual teamwork skills,” *Assessment & Evaluation in Higher Education*, vol. 42, no. 3, pp. 378–397, 2017.
34. E. Koh, H. Hong, and J. P.-L. Tan, “Formatively assessing teamwork in technology-enabled twenty-first century classrooms: exploratory findings of a teamwork awareness programme in singapore,” *Asia Pacific Journal of Education*, vol. 38, no. 1, pp. 129–144, 2018.
35. L. Riihelä, “Teaching information security: A systematic mapping study,” 2019. Master thesis.
36. M. Berson and I. Berson, “Bringing the cybersecurity challenge to the social studies classroom,” *Social education*, vol. 78, no. 2, pp. 96–100, 2014.
37. C. Turner and C. Turner, “Integrating cybersecurity into the sociology curriculum: the case of the password module,” *Journal of Computing Sciences in Colleges*, vol. 33, no. 1, pp. 109–117, 2017.
38. X. Mountrouidou and X. Li, “Cyber security education for liberal arts institutions,” in *Journal of The Colloquium for Information System Security Education*, vol. 6, pp. 17–17, 2019.
39. P. Bourque, R. E. Fairley, *et al.*, *Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0.* IEEE Computer Society Press, 2014.
40. A. A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne, and E. Motta, “The computer science ontology: a large-scale taxonomy of research areas,” in *International Semantic Web Conference*, pp. 187–205, Springer, 2018.
41. R. Martinez-Maldonado, A. Pardo, N. Mirriahi, K. Yacef, J. Kay, and A. Clayphan, “Latux: An iterative workflow for designing, validating, and deploying learning analytics visualizations.,” *Journal of Learning Analytics*, vol. 2, no. 3, pp. 9–39, 2015.
42. C. Izu, C. Schulte, A. Aggarwal, Q. Cutts, R. Duran, M. Gutica, B. Heinemann, E. Kraemer, V. Lonati, C. Mirolo, *et al.*, “Program comprehension: Identifying learning trajectories for novice programmers,” in *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education*, pp. 261–262, 2019.

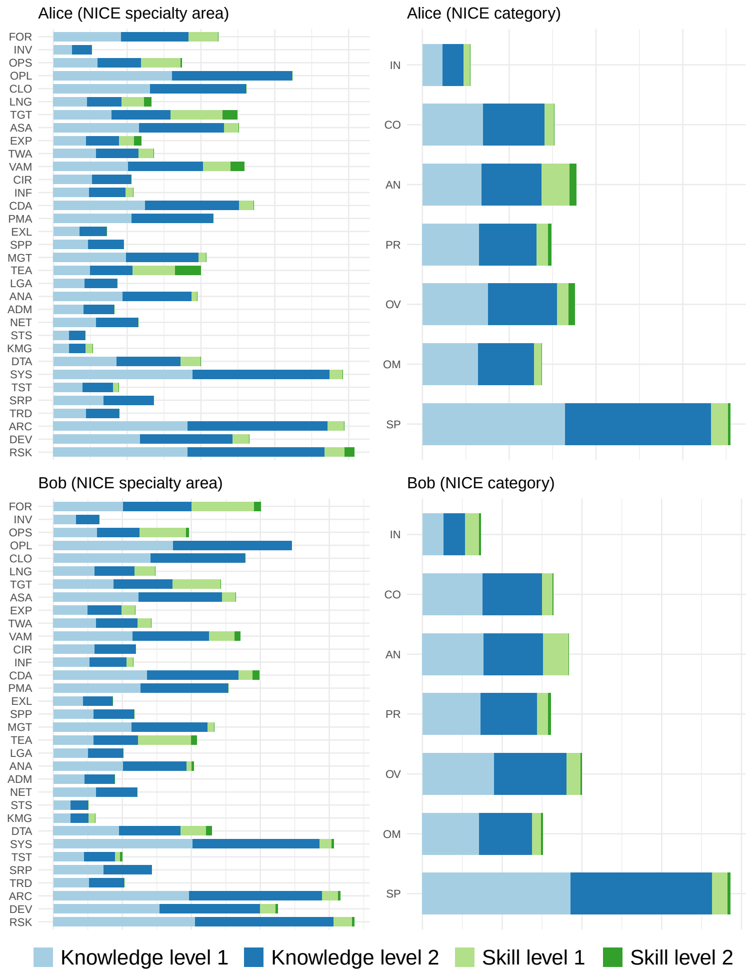


Fig. 2. Comparing the summarised results from two students using NICE framework specialty areas (on the left) and NICE framework high level categories (on the right). Please refer to NICE framework documentation [6] for detailed list of specialty areas, and categories.

Curriculum Vitae

Personal data

Name	Sten Mäses
Birth	1987, Tallinn, Estonia
Nationality	Estonian

Contact information

E-mail	mail@sten.ninja

Education

2016–2020	Tallinn University of Technology (TalTech) Computer Science, PhD
2011–2015	TalTech & University of Tartu Cyber Security, MSc *cum laude*

Languages

Estonian	native
English	fluent

Professional employment

2015– …	TalTech, early stage researcher & lecturer
2013–2015	ASA Quality Services, quality assurance specialist & trainer

Elulookirjeldus

Isikuandmed

Nimi	Sten Mäses
Sünd	1987, Tallinn, Eesti
Kodakondsus	Eesti

Kontaktandmed

E-post	mail@sten.ninja

Haridus

2016–2020	Tallinna Tehnikaülikool (TalTech)
	Informaatika, doktoriõpe
2011–2015	TalTech & Tartu Ülikool
	Küberkaitse, MSc *cum laude*

Keelteoskus

eesti keel	emakeel
inglise keel	kõrgtase

Teenistuskäik

2015– . . .	TalTech, nooremteadur & lektor
2013–2015	ASA Quality Services, kvaliteedispetsialist & koolitaja