Transdisciplinary perspectives on validity: bridging the gap between design and implementation for technology-enhanced learning systems

Haastrecht, M.A.N. van

TRANSDISCIPLINARY PERSPECTIVES ON VALIDITY

BRIDGING THE GAP BETWEEN DESIGN AND IMPLEMENTATION FOR TECHNOLOGY-ENHANCED LEARNING SYSTEMS

Max Anton Nicolaas van Haastrecht

Universiteit Leiden
2025

This document was typeset using LaTeX and has the typographical look-and-feel of the classicthesis package developed by André Miede and Ivo Pletikosić: https://bitbucket.org/amiede/classicthesis/.

ISBN: 978-94-6506-528-1
Printing: Ridderprint | https://ridderprint.nl

TRANSDISCIPLINARY PERSPECTIVES ON VALIDITY

BRIDGING THE GAP BETWEEN DESIGN AND IMPLEMENTATION FOR TECHNOLOGY-ENHANCED LEARNING SYSTEMS

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op vrijdag 24 januari 2025
klokke 13:00 uur

door

Max Anton Nicolaas van Haastrecht
geboren te Bloemendaal, Nederland
in 1995

PROMOTOR:
Prof.dr. M.R. Spruit

CO-PROMOTOR:
Dr. M.J.S. Brinkhuis (Universiteit Utrecht)

PROMOTIECOMMISSIE:
Prof.dr. M.M. Bonsangue
Prof.dr.ir. N. Mentens
Prof.dr.ir. J.M.W. Visser
Prof.dr. F. Dalpiaz (Universiteit Utrecht)
Prof.dr. S.A. Fricker (Fachhochschule Nordwestschweiz)
Prof.dr. A.G.J. van de Schoot (Universiteit Utrecht)

SIKS Dissertation Series No. 2025-01
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

The research of this dissertation was made possible through a European Commission Horizon 2020 grant for the GEIGER project (ID: 883588).

The things you regret most in life
are the things you didn't do.

— Steve Jobs

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Alle Sätze sind gleichwertig.
Der Sinn der Welt muß außerhalb ihrer liegen.

— Ludwig Wittgenstein

ACKNOWLEDGEMENTS

There are different ways to interpret the above statement by Wittgenstein, but I interpret it as a reminder that while propositions can describe facts about the world, they struggle to express deeper meaning or purpose. While the propositions in this dissertation have brought me to the point of defending my PhD, it is the people in my life who have given meaning and purpose to my PhD journey.

First of all, a massive thank you to Marco Spruit and Matthieu Brinkhuis, for your unwavering support throughout these years. Marco, thank you for giving me the chance to join your group in Utrecht and later in Leiden, and thank you for the trust you placed in me from day one. The way you allowed me to take on a leading role in GEIGER activities enabled me to grow as a researcher and as a person, and I honestly cannot think of many supervisors who would be as daring. Matthieu, you were always a cheerful and creative voice during our meetings and lunches. It is safe to say that the second half of this dissertation would have looked very different if you had not suggested looking into learning analytics and educational measurement for a change. It is even safer to say that I would not have my current job at Cito without your relentless enthusiasm and your support for me when looking for my next step. Thank you both for making these 4.5 years such a smooth ride.

That ride started in Utrecht in 2020, in the middle of the COVID pandemic. Jelmer Koorn, Ellen Deelen, Anna Wegmann, and Jan Posthoorn, thanks for helping me get settled during these strange times. Jan deserves a special mention for also having guided me through my MSc thesis at ORTEC, in a time when programming languages still seemed scarier to me than publication reviews.

In October 2021 I moved to Leiden University, where Bram van Dijk and Tom Kouwenhoven were so kind to welcome me. We organised the monthly LIACS PhD seminar together during my first year, which really helped me to get settled and get to know our amazing colleagues at LIACS. I want to thank Alexandra Blank for always being there to energise the PhD community at LIACS. You inspired me to organise the LIACS PhD/postdoc weekend, first with Nathan Schiele, then with Matthias Müller-Brockhausen. Nathan and Matthias, you have been an amazing asset to LIACS over the years, not in the least through your activism in fighting for the rights of Teaching PhDs.

From September 2022 onwards, I spent two years as a representative for PhDoc in Leiden's University Council. Thank you Lis Kerr and Niko Kontovas for all the work you did and for granting me this opportunity, and good luck to Tahmina Fariaby and Marie Kolbenstetter in the coming years. There are far too many people that I met during my time in the council to thank them all, but thank you to everyone who is going above and beyond to make Leiden University an inspiring home for students and staff. I do want to mention a few people specifically. Janneke Vader and Anita Romijn, thank you for making Leiden a better place for PhDs and postdocs. Pauline Vincenten and Robert-Jan van Ette, thank you for all the work you are doing for our university. Leiden is lucky to have you.

Thank you to all those who have been a part of Marco's TDS Lab over the years, who have welcomed me in Utrecht and later made me feel at home in Leiden. Armel Lefebvre, Chaïm van Toledo, Emil Rijcken, Friso van Dijk, Hielke Muizelaar, Jim Achterberg, Marcel Haas, Noha Tawfik, Pablo Mosteiro Romero, Samar Samir, whether you know it or not, you have all inspired me in your own way.

Bilge Yigit Ozkan, thank you for your support during my first year in Utrecht. You helped me discover what it means to do research. Alireza Shojaifar, thank you for being a great collaborator and for granting me the honour to be a paranymph at your defence. Injy Sarhan, thank you for being a wonderful friend and colleague, and for all the work you did in the GEIGER project, making me look a lot smarter and productive than I really am. I wish you all the best in life, you deserve it.

Maarten Hamming and Bram van Dijk, thank you for supporting me as paranymphs on my final day in Leiden. You seem a fitting duo, with Bram being the longest-standing colleague and friend from my time in Leiden, and Maarten being the longest-standing friend from my time in Groningen.

Finally, a massive thank you to the people closest to me, who were so kind to put in the effort to understand what I was doing, even when I did not always understand it myself. Mom, dad, Thomas, and Ties, thank you for being who you are and for making me who I am. Veerle, I could not have done this without you, and would not have wanted to either. Everything just seems so much easier when I am with you.

1

INTRODUCTION

The contents of this chapter are based on: Van Haastrecht (2021). Doctoral Consortium. European Conference on Information Systems.

You're a hairdresser with a small salon tucked away somewhere in the Swiss countryside. You have a fixed group of returning customers that you know well, to the point that you consider them to be trusted friends. One of your elderly customers has just sent you an e-mail asking if you can help with a document they cannot open for some reason. You hesitate for a moment, since you're not particularly tech-savvy yourself. You decide it's worth a try, manage to open the document, but find that it's empty. You click around to see if anything happens. Your computer responds strangely for a few seconds. Then everything goes back to normal. You conclude that whoever shared this document with your elderly customer must have made a mistake.

A few days later, the elderly customer drops in for an appointment. You tell them about your finding, but they seem confused. They haven't e-mailed you recently. Now you're the one who's confused, and you check the e-mail you received. All of a sudden you notice that the e-mail address looks similar to the customer's, but is definitely different. You start to get scared and then the phone rings. It's the bank. Someone on the other side of the world has just spent thousands in a casino and your account is blocked. The money is gone. You have a dejected look on your face and the customer asks what's wrong. Nothing, you say.

Although it may seem dramatised, some version of this story occurs on a daily basis at small businesses across Europe and the world. Small- and medium-sized enterprises (SMEs) make up 99% of all companies in the EU (European Commission, 2016). SMEs are more vulnerable to cyber threats than larger companies, due to their limited cybersecurity knowledge and resources (Heidt et al., 2019). This makes them an ideal target for cybercriminals. A 2019 report surveying 2,176 small businesses showed that 66% experienced a cyberattack in the preceding 12 months (Ponemon Institute, 2019).

The GEIGER project (GEIGER Consortium, 2020) aimed to address the cybersecurity challenge faced by SMEs by providing a trusted solution for assessing cybersecurity risk. The work of this dissertation centres around the activities of the GEIGER project. In this introduction, we will cover why projects like GEIGER are necessary to solve the cybersecurity challenge SMEs face, how we approached the process of finding a solution for this challenge, and what methods we used to find answers to concrete research questions about this challenge.

1.1 WHY

We established in the previous paragraphs that SMEs tend to lack the cybersecurity knowledge and resources required to deal with the cyber attacks they regularly face. Given that SMEs comprise 99% of all businesses in the EU, it is no wonder that the European Commission is intent on helping these businesses to protect themselves.

However, protecting SMEs against cyber threats is not trivial. Although cybersecurity is often primarily seen as a technical challenge, it is the human element that regularly forms the weakest link at SMEs (Shojaifar, Fricker, and Gwerder, 2020). A project like GEIGER, therefore, should not only offer technical countermeasures to cyber threats, but should also educate SME employees to increase cybersecurity awareness. In fact, one could even argue that having a basic level of awareness about the existence of cyber threats is a prerequisite for an SME to be motivated to protect themselves. The GEIGER project attempted to solve this apparent catch-22, where awareness is a prerequisite for motivation and motivation is a prerequisite for awareness, by fostering trust.

We know from self-determination theory (SDT) (Deci and Ryan, 1985; Ryan and Deci, 2000) that the psychological needs of autonomy, competence, and relatedness are what drive motivation. We realised early on in the GEIGER project that perceived autonomy and competence were difficult to influence, as SMEs often do not yet have the cybersecurity knowledge to act independently and effectively. This leaves perceived relatedness as the primary need which can be externally influenced.

Consider again the example of the hairdresser in the opening paragraphs of this introduction. The hairdresser regards their customers to be trusted

friends. Supposing one of these trusted friends would have followed a training to become a cybersecurity expert, this friend could then motivate the hairdresser to improve the cybersecurity maturity of the salon by appealing to the connection and mutual trust they have. Training trusted advisors to become security defenders creates a pathway towards motivating SMEs to become more secure. The strategy to actively involve trusted security defenders is unique to the GEIGER project, and we now have a sense of why such an approach is necessary for a solution to the cybersecurity challenges faced by SMEs.

Nevertheless, the socio-technical context of SMEs is complex, and training a trusted security defender is just a small piece in the overall GEIGER puzzle. Figure 1.1 shows the full ecosystem of the GEIGER project, highlighting both its social and technical elements. Associations and networks provide information to SMEs regarding the GEIGER application. Security defenders act as a trusted advisor and help SMEs to make a smooth start with installation and taking the first steps. GEIGER helps SME employees to become more aware of cybersecurity topics, as well as helping the business to assess and manage their cybersecurity risks. Cybersecurity tool and service providers contribute technical countermeasures that SMEs can implement, while Computer Emergency Response Teams (CERTs) provide information on the threat landscape that can be used to prioritise threats for users and to issue notifications. In the ideal situation, the SME improves their cybersecurity awareness while countering technical security risks, with tech-savvy employees potentially becoming the new generation of security defenders. The SME can thus itself contribute to making future businesses more secure.

Figure 1.1 gives a sense of the complexity of helping SMEs improve their cybersecurity. We need to find a balance between a solution that is technically sound and designed based on rigorous principles, while concurrently ensuring that users with relatively little knowledge about the topic of cybersecurity stay motivated and engaged. The frequently conflicting values of rigour and simplicity in a socio-technical context are characteristic to the class of *wicked problems* (Buchanan, 1992; Rittel and Webber, 1973). Rittel and Webber (1973) provide some further properties of such problems, which include: "there is no definitive formulation of a wicked problem", "wicked problems have no stopping rule", and "solutions to wicked problems are not true-or-false, but good-or-bad." All of these properties apply to our context of SME cybersecurity risk assessment. There is no definitive way to formulate and approach SME cybersecurity risk assessment. There is no such thing as absolute security and, therefore, no stopping rule stating that SMEs have done all they can to counter cybersecurity threats. Finally, there is not a single right way to assist SMEs, but rather a whole spectrum of strategies, where one strategy may focus primarily on the social elements of the socio-technical system and another may focus primarily on technical elements. Strategies may be poorly implemented or unsuccessful, but cannot be a priori false. We

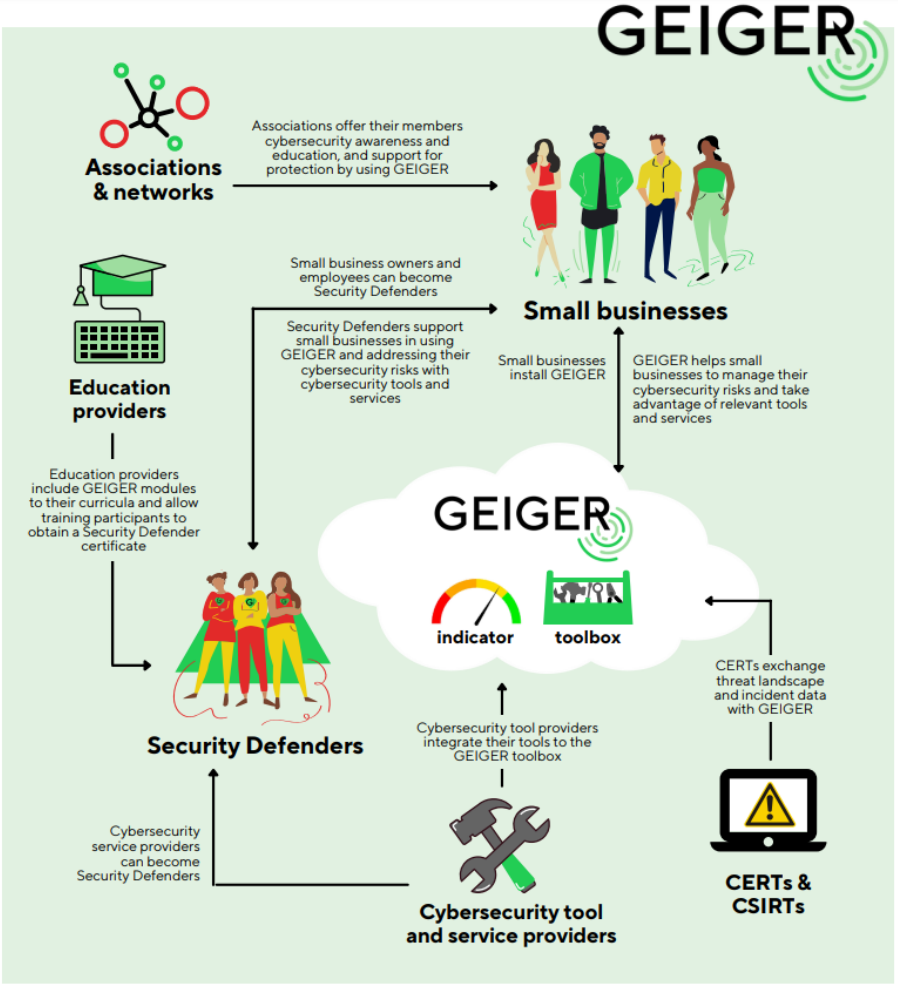


Figure 1.1: The socio-technical ecosystem of the GEIGER project. Used with permission from the creator of the visualisation, Heini Järvinen. Source: `https://cyber-geiger.eu/`.

detail the research strategy we use for this dissertation, the how, in the next section.

1.2 HOW

Our overarching research methodology should be suited to the socio-technical, complex, and wicked nature of our cybersecurity problem and should accommodate the integration of knowledge from several different research fields, such as cybersecurity and education. Additionally, our methodology needs to facilitate the active involvement of all stakeholders, including voices from academia and society. Transdisciplinary research is a research strategy that addresses our requirements exceptionally well.

Jantsch (1970) originally defined transdisciplinary research as: "the coordination of all disciplines and interdisciplines in the education/innovation system." Over time, the concept of transdisciplinarity evolved to explicitly include societal partners beyond the education system, and to aim at performing societally relevant research through reflexive practice (Lawrence et al., 2022). Figure 1.2 depicts how transdisciplinary research differs from traditional strategies such as disciplinary, participatory, and interdisciplinary research. By crossing both disciplinary and sectoral boundaries, transdisciplinary research stimulates the development of integrated knowledge that benefits both science and society.

Lawrence et al. (2022) outline three phases of the transdisciplinary research process. The first phase involves framing the research problem, the second phase involves the co-creation of transferable knowledge by societal and academic actors, and the third phase aims to integrate and apply the newly created knowledge. Lawrence et al. (2022) stress that "often the whole sequence or individual phases need to be iterated, and the phases often run in parallel." This is another reminder that wicked, complex problems call for solutions that are themselves rather complex. An issue that arises with the transdisciplinary research process is that although it helps to describe how we will tackle our overarching research problem, it gives minimal guidance on the exact research questions that should be answered and the research methods that could be used.

To bridge the gap between the why and the how, we use the engineering cycle of Wieringa (2014). In design science, the cyclic process of design generally includes phases of problem framing, design, and evaluation. Wieringa refers to these phases as problem investigation, treatment design, and treatment validation. However, Wieringa extends the design cycle with a fourth phase of treatment implementation: "the application of the treatment to the original problem context." Where design science research projects are generally concerned with the first three phases, our work in the GEIGER project had the express intent of applying the designed solution within the original problem context. The engineering cycle therefore offers a better fit to our re-

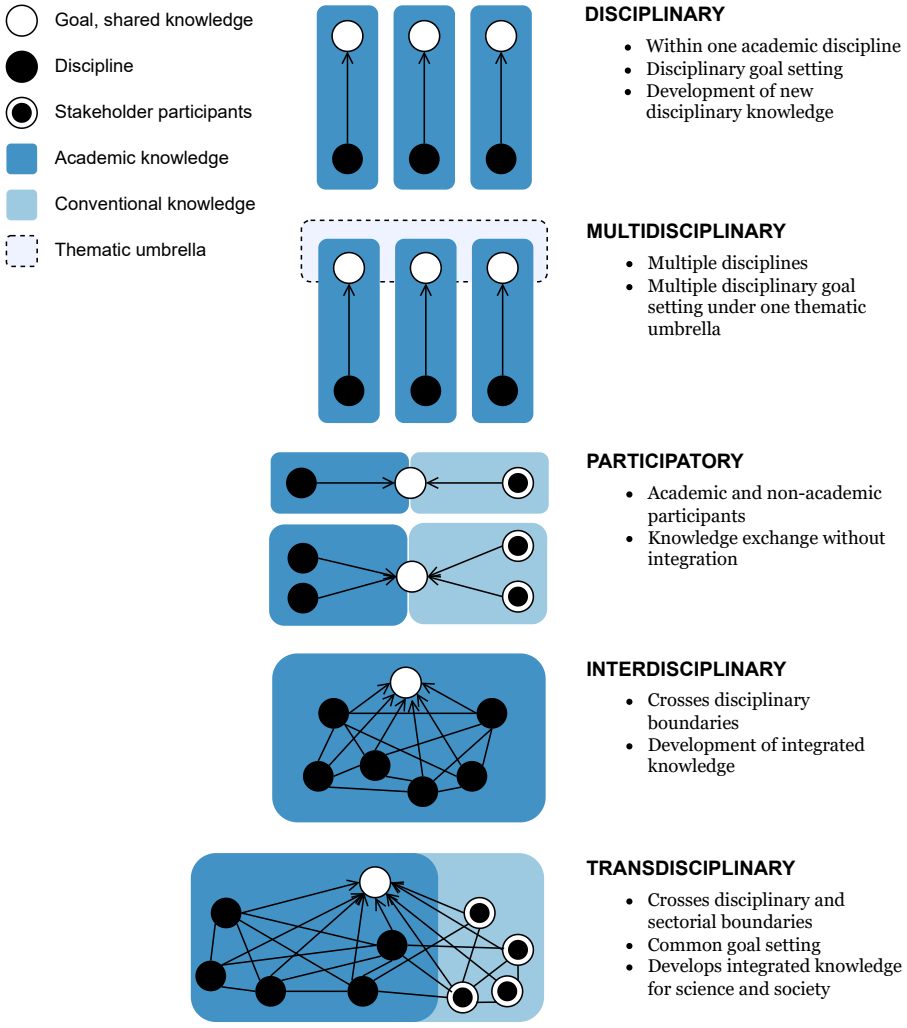


Figure 1.2: Comparison of transdisciplinary research to more traditional research strategies. This visualisation is based on Morton et al. (2015) and Tress et al. (2005), with a difference being that we consider participatory research to involve stakeholder participants.

Table 1.1: The four parts of this dissertation, with the corresponding phases of the transdisciplinary research process and the engineering cycle indicated.

PART	TRANSDISCIPLINARY RESEARCH PROCESS PHASE	ENGINEERING CYCLE PHASE
I	Problem framing	Problem investigation
II	Co-creation	Treatment design
III	Co-creation & integration and application	Treatment validation
IV	Integration and application	Treatment implementation

search project than the design cycle, and the treatment implementation phase aligns well with the integration and application phase of the transdisciplinary research process.

Perhaps most importantly, Wieringa's engineering cycle suggests concrete knowledge questions and design problems that are paired to each phase in the cycle. During the first phase of problem investigation, Wieringa suggests to address knowledge questions regarding the involved stakeholders, the conceptual problem framework, and the phenomena that arise in the problem setting. A research method suggested by Wieringa for the problem investigation phase is a survey, or systematic review. The second phase, treatment design, involves specifying requirements, surveying available treatments, and designing new treatments. In the GEIGER setting, this involves collecting user requirements from SMEs and incorporating these requirements into a newly designed cybersecurity risk assessment application. The treatment design phase, therefore, involves research methods centred around collaborative design together with stakeholders and use case experiments to demonstrate the viability of developed artefacts.

The third phase of the engineering cycle concerns treatment validation. Wieringa suggests to address the knowledge questions of this phase, regarding whether our new designs produce the intended effects, using action research methods supplemented with techniques to infer information from data, such as grounded theory. In the fourth and final phase of treatment implementation, we aim to answer questions concerning the implemented artefact, such as to what extent the artefact contributes to stakeholder goals. In the GEIGER setting, this could involve questionnaires aimed at SME users, but could also involve interviews with educational technology experts regarding the ability of a solution like GEIGER to contribute to the educational experience of SMEs. Table 1.1 provides an overview of how the different parts of this dissertation correspond to the phases of the engineering cycle and the transdisciplinary research process.

Figure 1.3 visualises the connection between the transdisciplinary research process and the engineering cycle, and connects the topics of our chapters to the respective phases of both. The research methods used in the chapters are informed by the research methods suggested by Wieringa for the various phases of the engineering cycle. We additionally show how we gradually

included knowledge from different scientific disciplines and non-academic stakeholders to evolve from a simple interdisciplinary setting to a true transdisciplinary project.

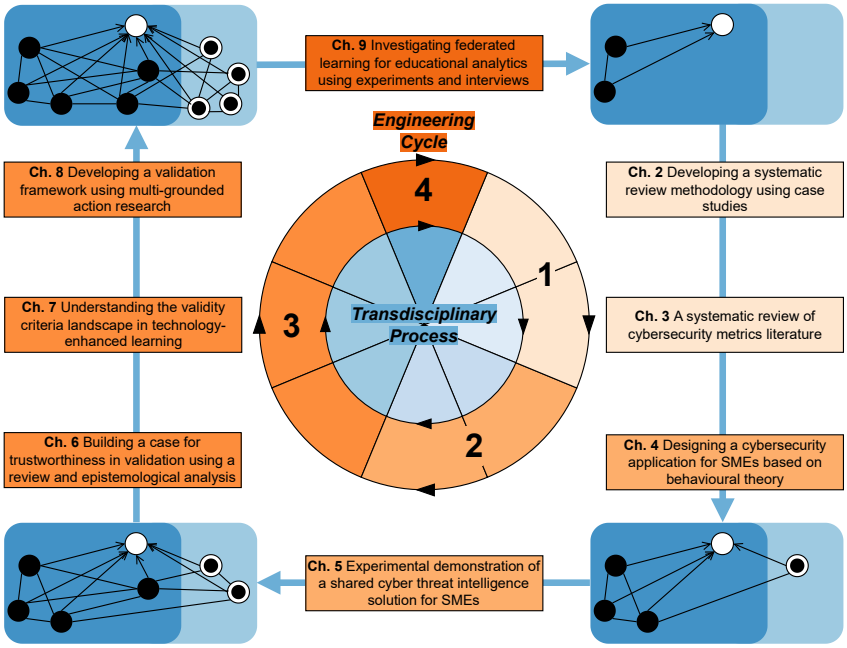


Figure 1.3: A visualisation of our research process. We combine the transdisciplinary process described by Lawrence et al. (2022) and the engineering cycle of Wieringa (2014). An overview of the different phases of the transdisciplinary research process and the engineering cycle is provided in Table 1.1.

Recall that the transdisciplinary research process and the engineering cycle emphasise that there is no true end to the research process, just as there is no stopping rule for wicked problems. Rather, a first cycle of the research process generates new ideas and hypotheses for the next cycle. In our concluding Chapter 10, we will reflect on possibilities for future research cycles. For now, we will turn our attention to the methods we intend to use to find answers to concrete research questions regarding the challenge of using a technology-enhanced learning (TEL) solution to educate and assess SMEs on the topic of cybersecurity.

1.3 WHAT

Inspired by the goals of the GEIGER project, the main research question of this dissertation is:

How can transdisciplinary research inform the design and validation of technology-enhanced learning solutions?

In the following paragraphs, we will cover the various sub-questions that are addressed in the chapters of this dissertation. The chapters and questions are ordered using the phases of the transdisciplinary research process and the engineering cycle.

PART I of this dissertation covers the problem investigation phase of the engineering cycle, and consists of Chapter 2 and Chapter 3.

CHAPTER 2 addresses the question: What are the elements of an accessible and swift systematic review methodology? We begin our research with the problem framing phase of the transdisciplinary process and the problem investigation phase of the engineering cycle. Systematic literature reviews are commonly used to create an overview of existing literature in a specific research domain. However, systematic reviews are time-intensive affairs and traditional approaches that rely purely on database searches regularly leave out grey literature such as technical reports. In a field such as cybersecurity, where reports from industry are a common source of knowledge, traditional systematic review methodologies can thus be problematic. This provided the motivation to develop a novel systematic review methodology, SYMBALS, that incorporates active learning innovations to speed up the process and a snowballing phase to better cover grey literature. We use two case studies to demonstrate the effectiveness of this method.

CHAPTER 3 addresses the question: How can SME cybersecurity be measured? Using our novel systematic review methodology SYMBALS, we conduct a systematic review of cybersecurity metrics literature, to gain insight into how cybersecurity indicators are measured in the complex socio-technical context of SMEs. This chapter is part of the problem framing and problem investigation phases, as it helps to answer questions regarding the conceptual framework that we can employ in the design phase that follows. The key artefact produced is a socio-technical cybersecurity framework for SMEs that contains insights relevant to practice.

PART II of this dissertation covers the treatment design phase of the engineering cycle, and consists of Chapter 4 and Chapter 5. We combine our insights from the problem investigation phase with elicited user requirements, to design a relevant solution with a rigorous foundation.

CHAPTER 4 addresses the question: How should an SME cybersecurity application be designed to motivate users? This chapter therefore moves from the introductory problem framing and investigation phases to the phases

related to co-creation and design. Through a collaborative design research approach, we design a first version of our cybersecurity application based on insights from behavioural theories. The presented design is the result of an iterative process of eliciting SME user requirements and feedback to inform design improvements. We contribute to societal knowledge in two ways. Firstly, through the direct interaction with SME stakeholders in the GEIGER project. Secondly, via the dissemination of our cybersecurity risk assessment application to the broader public, the resulting artefact contributes to our understanding of how ideas from behavioural theories can be used to guide design choices.

CHAPTER 5 addresses the question: How can cyber threat intelligence be incorporated in an SME cybersecurity application? In collaboration with the Romanian CERT, we develop a shared cyber threat intelligence platform, and demonstrate the ability of the GEIGER application to turn advanced cyber threat intelligence into actionable suggestions for SMEs. The research performed in this chapter can be described as technical action research, which Wieringa (2014) defines as "the use of an artefact prototype in a real-world problem to help a client and to learn from this." The artefact prototype is our threat intelligence platform, and the client is the SME user. Key contributions are a detailed process description of how threat intelligence can be turned into actionable insights, and a bolstering of societal knowledge through the co-creation of the platform with industry partners.

PART III of this dissertation covers the treatment validation phase of the engineering cycle, and consists of Chapter 6, Chapter 7, and Chapter 8. Besides shifting the focus from design to validation, this part of the dissertation additionally shifts from a narrow, context-specific view used to design an educational cybersecurity application for SMEs (GEIGER), to a broad view used to develop a validation framework for TEL more generally. GEIGER is an example of a TEL application, where analytics regarding SME employee performance in various cybersecurity learning activities are used to inform an eventual SME cybsercurity risk assessment. To holistically validate the GEIGER solution, we thus need a holistic validation framework for TEL solutions. Part III aims to develop such a framework.

CHAPTER 6 addresses the question: Which criteria are essential to a holistic validation strategy for an educational application? Chapter 6 moves us into the treatment validation phase, where we ask questions about how we can assess the effectiveness of our designed artefact. In terms of the transdisciplinary research process, we are balancing between the co-creation and integration phases. We are both in the process of co-creating knowledge about our designed artefact and reflecting on what is required to create impact in science and society with our final solution. In this chapter, we theorise about

the epistemological basis required for validity considerations in learning analytics. By conducting a systematic review of learning analytics validation approaches, we create an overview of how existing validity criteria are used in a Learning Analytics Validation Assistant (LAVA), which can aid researchers in developing holistic validation strategies.

CHAPTER 7 addresses the question: How are validity criteria applied in TEL research? This chapter is part of the same engineering cycle and transdisciplinary research process phases as Chapter 6, and can be considered an extension of that work. We conduct a systematic literature review using SYMBALS, to uncover which validity criteria are considered in TEL research, which methods are used to gain insight into these criteria, and whether they are on average assessed positively or negatively. By comparing validity criteria definitions and usage over time, we create a picture of the validity criteria landscape, which can inform future holistic validation frameworks.

CHAPTER 8 addresses the question: How can e-assessment solutions be validated comprehensively and practically? We employ a multi-grounded action research (Goldkuhl, Cronholm, and Lind, 2020; Karlsson and Ågerfalk, 2007) approach to develop a validation framework for e-assessment solutions such as GEIGER. Multi-grounded action research contains elements of grounded theory and action research, and is therefore suited to the treatment validation phase of the engineering cycle and to transdisciplinary theorising. As with the previous two chapters, the research of this chapter sits in the balance of the co-creation and integration phases of the transdisciplinary process. Since our validation framework is developed with repeated, active input from project partners, it not only contributes to the scientific literature, but also introduces societal stakeholders to valuable insights concerning validation strategies.

PART IV of this dissertation covers the treatment implementation phase of the engineering cycle, and consists of Chapter 9. We reflect on the question of what happens after a validated solution is implemented in practice. In our concluding Chapter 10, we look ahead to which research hypotheses could be addressed in a next iteration of our engineering cycle.

CHAPTER 9 addresses the question: How does the privacy-performance trade-off manifest itself in educational analytics? We conduct technical experiments to demonstrate the potential of privacy-preserving machine learning in an educational analytics context. Through the preliminary results of a series of interviews with educational technology experts, we reflect on the viability of introducing advanced machine learning techniques into educational contexts. This mixed-methods study brings to light several conditions for a successful implementation of an educational innovation such as the GEIGER application, and can therefore be considered as part of the treatment implementation phase

Table 1.2: An overview of the main research question and sub-questions addressed in this dissertation. We indicate how the individual studies relate to the transdisciplinary process and the engineering cycle. Additionally, we specify the artefacts resulting from our studies.

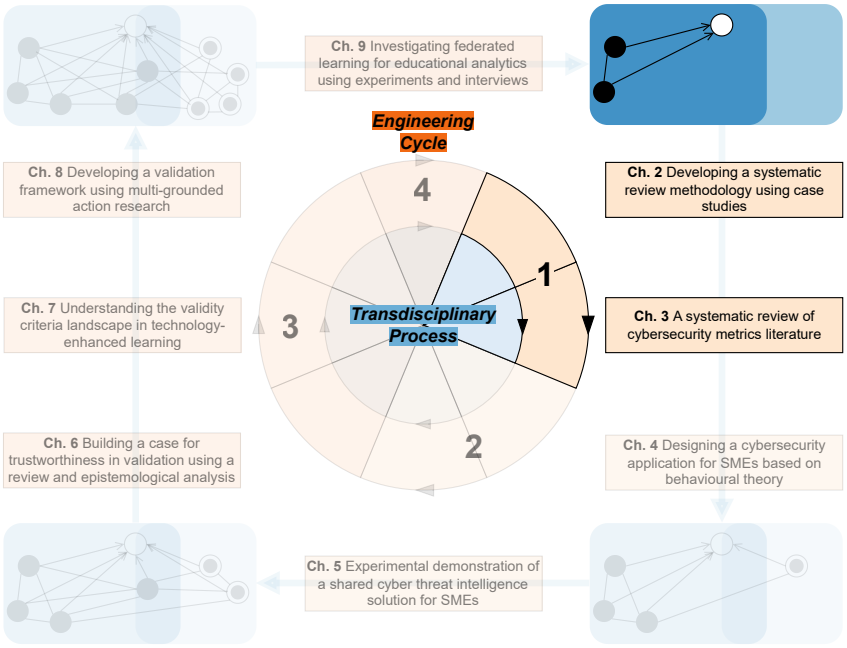
CH.	RESEARCH QUESTION	PROCESS	CYCLE	ARTEFACT
Main	How can transdisciplinary research inform the design and validation of technology-enhanced learning solutions?	-	-	-
2	RQ Ch. 2 What are the elements of an accessible and swift systematic review methodology?	Problem framing	Problem investigation	SYMBALS
3	RQ Ch. 3 How can SME cybersecurity be measured?	Problem framing	Problem investigation	Socio-technical cybersecurity framework
4	RQ Ch. 4 How should an SME cybersecurity application be designed to motivate users?	Co-creation	Treatment design	SME cybersecurity algorithm
5	RQ Ch. 5 How can cyber threat intelligence be incorporated in an SME cybersecurity application?	Co-creation	Treatment design	Cyber threat intelligence platform
6	RQ Ch. 6 Which criteria are essential to a holistic validation strategy for an educational application?	Co-creation, integration and application	Treatment validation	LAVA
7	RQ Ch. 7 How are validity criteria applied in technology-enhanced learning research?	Co-creation, integration and application	Treatment validation	Validity criteria landscape
8	RQ Ch. 8 How can e-assessment solutions be validated comprehensively and practically?	Co-creation, integration and application	Treatment validation	VAST
9	RQ Ch. 9 How does the privacy-performance trade-off manifest itself in educational analytics?	Integration and application	Treatment implementation	FLAME

of the engineering cycle and the integration phase of the transdisciplinary process.

CHAPTER 10 , finally, reflects on the findings of the previous chapters. Using the insights we gained, we consider the possibilities for future research cycle iterations. Table 1.2 summarises the research questions of this dissertation, indicating their positions in the transdisciplinary research process and the engineering cycle. Not every individual chapter explicitly contributes knowledge to science and society, but the sum of all individual parts possesses the clear characteristics of a transdisciplinary research project. In Chapter 10, we will discuss whether our transdisciplinary approach has been successful in tackling our wicked problem.

Part I

PROBLEM INVESTIGATION



2

SYMBALS: A SYSTEMATIC REVIEW METHODOLOGY

Research output has grown significantly in recent years, often making it difficult to see the forest for the trees. Systematic reviews are the natural scientific tool to provide clarity in these situations. However, they are protracted processes that require expertise to execute. These are problematic characteristics in a constantly changing environment. To solve these challenges, we introduce an innovative systematic review methodology: SYMBALS. SYMBALS blends the traditional method of backward snowballing with the machine learning method of active learning. We applied our methodology in a case study, demonstrating its ability to swiftly yield broad research coverage. We proved the validity of our method using a replication study, where SYMBALS was shown to accelerate title and abstract screening by a factor of 6. Additionally, four benchmarking experiments demonstrated the ability of our methodology to outperform the state-of-the-art systematic review methodology FAST^2.

The contents of this chapter are based on: van Haastrecht, Sarhan, Yigit Ozkan, et al. (2021). SYMBALS: A systematic review methodology blending active learning and snowballing. Frontiers in research metrics and analytics.

2.1 INTRODUCTION

Both the number of publishing scientists and the number of publications are constantly growing (Ware and Mabe, 2015). The natural scientific tool to provide clarity in these situations is the systematic review (Glass, 1976), which has spread from its origins in medicine to become prevalent in a wide number of research areas (Petticrew, 2001). Systematic reviews offer a structured and clear path to work from a body of research to an understanding of its findings and implications (Gough et al., 2017; Higgins et al., 2019). Systematic reviews are ubiquitous in today's research. A search in the Scopus abstract database for the phrase 'systematic review' yields more than 45,000 results for the year 2020 alone.

Nevertheless, systematic reviews have shortcomings. They are particularly protracted processes (Borah et al., 2017; O'Connor et al., 2019), that often require an impractical level of expertise to execute (Zhang and Ali Babar, 2013). These issues have been recognised for decades (Petticrew, 2001), but not solved. This hampers our ability as researchers to apply this potent tool in times where change is ceaseless and sweeping.

However, with recent advances in machine learning and active learning, new avenues for systematic review methodologies have appeared (Marshall and Wallace, 2019). This is not to say that these techniques make traditional systematic review techniques obsolete. Methodologies employing automation techniques based on machine learning are often found to omit around 5% of relevant papers (Gates et al., 2019; Yu, Kraft, et al., 2018; Yu and Menzies, 2019). Additionally, usability and accessibility of automation tools is a common issue (Gates et al., 2019; Harrison et al., 2020) and many researchers do not trust machine learning methods enough to fully rely on them for systematic reviews (O'Connor et al., 2019).

Therefore, in this chapter, we argue for the combination of the proven method of backward snowballing (Wohlin, 2014) with novel additions based on machine learning techniques (van de Schoot et al., 2021). This yields SYMBALS: a SYstematic review Methodology Blending Active Learning and Snowballing. The challenges faced by systematic review methodologies motivate the research question of this chapter:

- **RQ**: How can active learning and snowballing be combined to create an accessible and swift systematic review methodology?

The remainder of this chapter is structured as follows. In Section 2.2, we cover related work on systematic review methodologies and active learning techniques for systematic reviews. In Section 2.3, we introduce SYMBALS, our innovative systematic review methodology. We explain each step of the methodology in detail. Section 2.4 evaluates and demonstrates the effectiveness of our methodology using two case studies: a full application of SYMBALS 2.4.1 and a benchmarking study 2.4.2. In Section 2.5, we discuss

the implications of the case studies and the limitations of our research. Finally, we conclude and present ideas for future research in Section 2.6.

2.2 RELATED WORK

2.2.1 *Systematic review methodologies*

From its origins (Glass, 1976) and main application in the field of medicine, the use of systematic reviews has spread across the research community (Petticrew, 2001). In the area of information systems, the use of this tool was limited only two decades ago (Webster and Watson, 2002). Yet, systematic reviews are ubiquitous in the field now.

Software engineering is a field of research that has been specifically active in propelling systematic review practice. Since the first push for Evidence-Based Software Engineering (EBSE, (Kitchenham, Dyba, et al., 2004)), many contributions to systematic review practice have been made. Learning from applying the process in their domain (Brereton et al., 2007), clear guidelines for performing systematic reviews were developed (Kitchenham and Charters, 2007). These guidelines have been implemented and new methodologies have been developed and formalised. An example is the snowballing methodology (Wohlin, 2014).

Hybrid strategies have emerged which combine results from abstract databases with snowballing (Mourão, Kalinowski, et al., 2017; Mourão, Pimentel, et al., 2020), as well as those that suggest automating certain steps of the systematic review process with machine learning techniques (Osborne et al., 2019). The use of systematic reviews in software engineering has matured to a stage where even tertiary studies - reviews of reviews - are common (Kitchenham, Pretorius, et al., 2010). These studies focus on issues such as orientation towards practice (F. Q. B. da Silva et al., 2011), quality evaluation (Khan et al., 2019), and time investment (Zhang and Ali Babar, 2013). Tertiary studies give insight into what constitutes a high-quality systematic review. We used these insights in constructing our methodology.

Even with all of the developments in systematic review methodologies, challenges remain. At the heart of these challenges lie the tradeoffs between automation and completeness and between automation and usability. Approaches using automation techniques to speed up the systematic review process generally miss approximately 5% of the relevant papers that would have otherwise been found (Gates et al., 2019; Yu, Kraft, et al., 2018; Yu and Menzies, 2019). Additionally, many automation tools for systematic reviews still suffer from usability issues. Some tools are evaluated as hard to use (Gates et al., 2019), while others are not suitable due to limited accessibility (Harrison et al., 2020).

The usability issues are certainly solvable. Certain automation tools already offer a good user experience (Harrison et al., 2020) and some are making their

code available open-source (van de Schoot et al., 2021), making these tools increasingly accessible and transparent. The concerns regarding completeness remain. However, we should be aware that the metric used to assess completeness - the percentage of the total relevant papers found using an automated process (Gates et al., 2019) - is quite strict. The metric assumes that the complete set of relevant papers were found in the original review, meaning the automated method can at best perform equally well.

With SYMBALS we advocate for the adoption of usable and accessible automation tools, specifically those facilitating active learning for title and abstract screening. By combining automation with backward snowballing, we hope to address the completeness concerns that are still prevalent in many fully automated methods. Given the relative novelty and complexity of active learning techniques, we opt to provide further explanation and contextualisation of active learning in Section 2.2.2.

2.2.2 *Active learning for systematic reviews*

Active learning is a machine learning method whereby a learning algorithm chooses the most relevant data points to learn from. The key concept motivating this approach is that the algorithm will perform better with fewer training samples if it can guide the learning process towards the most informative samples (Settles, 2012). This makes it very well suited to be applied in the title and abstract screening phase of systematic reviews, where researchers often start with a large set of papers and prefer to not perform the full time-consuming task manually (Yu, Kraft, et al., 2018).

Active learning for title and abstract screening works as follows. Researchers construct a dataset of potentially relevant research, with at least a title and abstract for each paper. Researchers should then define an initiation process and an appropriate stopping criterion for the active learning algorithm. The exact initiation process will differ, but the initial sample provided to the algorithm should contain at least one relevant and one irrelevant paper for the algorithm to learn from. At the same time, the sample should be relatively small compared to the complete set of papers, as there is no time advantage in this phase of the process.

After the algorithm has learned from the initial samples, it will present the researchers with the most informative paper first (Yu and Menzies, 2019). The researcher indicates whether the paper is relevant or irrelevant and the algorithm uses this input to retrain. The key challenge is to balance exploration and exploitation. The algorithm should learn to distinguish relevant from irrelevant papers as quickly as possible (exploration) while presenting the researchers with as many relevant papers as possible (exploitation). Active learning techniques have been shown to significantly reduce the time spent on title and abstract screening (Miwa et al., 2014), while minimally affecting the total number of relevant papers found (Yu, Kraft, et al., 2018). Using active

learning for title and abstract screening can intuitively be characterised as "researcher-in-the-loop" (van de Schoot et al., 2021) machine learning. Figure 2.1 depicts the active learning process using Business Process Model and Notation (BPMN).

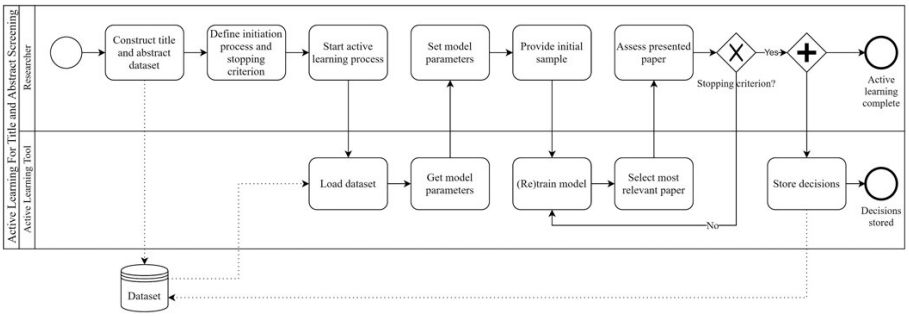


Figure 2.1: The active learning for title and abstract screening process, depicted using BPMN. One can clearly see why this process is characterised as "researcher-in-the-loop" (van de Schoot et al., 2021) machine learning.

In an evaluation of 15 software tools that support the screening of titles and abstracts (Harrison et al., 2020), Abstrackr (Wallace et al., 2012), Covidence (Babineau, 2014), and Rayyan (Ouzzani et al., 2016) emerged as the tools that scored best. FASTREAD (Yu, Kraft, et al., 2018) and ASReview (van de Schoot et al., 2021) are two additional tools incorporating active learning that have recently been introduced.

The first research using active learning techniques to supplement systematic reviews is beginning to appear. For the steps of 'identify research' and 'select studies' (Kitchenham, Budgen, et al., 2015), some suggest using active learning on database results as the sole method (Yu and Menzies, 2019). This yields a fast approach, as seen with the FASTREAD (Yu, Kraft, et al., 2018) and FAST^2 (Yu and Menzies, 2019) methodologies. However, these methods sacrifice a degree of completeness to manual screening (Gates et al., 2019), which itself can omit up to 30% of the relevant papers that could have been found by additionally using other techniques than database search (Mourão, Kalinowski, et al., 2017; Mourão, Pimentel, et al., 2020).

Approaches relying solely on database search also have no way of incorporating grey literature. Grey literature is research that does not originate from traditional academic publishing sources, such as technical reports and dissertations. This issue could be solved by searching for grey literature before screening (Rios et al., 2020), although this requires the researchers to know where to find relevant grey sources. The issues relating to the completeness of the review can be solved by incorporating a backward snowballing phase after database searching and screening (Mourão, Kalinowski, et al., 2017; Mourão, Pimentel, et al., 2020), which is exactly what we suggest to do in our approach.

Table 2.1: Overview of systematic review methodologies discussed in Section 2.2, the methods they use, and the properties they possess.

	Methods			Properties	
RESEARCH	DB SEARCH	AUTOMATION	SNOWBALLING	ACCESSIBLE	SWIFT
SYMBALS	✓	✓	✓	✓	✓
Miwa et al. (2014)	✓	✓	×	×	✓
Wohlin (2014)	×	×	✓	✓	×
Ros et al. (2017)	✓	✓	✓	×	✓
Mourão, Kalinowski, et al. (2017)	✓	×	✓	✓	×
Yu, Barik, et al. (2018)	✓	✓	×	×	✓
Yu and Menzies (2019)	✓	✓	×	×	✓
Mourão, Pimentel, et al. (2020)	✓	×	✓	×	✓
Rios et al. (2020)	✓	✓	×	×	✓

Active learning is not the only machine learning approach used to automate systematic reviews. Some researchers have suggested using natural language processing techniques to aid database search (Marcos-Pablos and García-Peñalvo, 2020; Osborne et al., 2019), while others prefer to use reinforcement learning in title and abstract screening, rather than active learning (Ros et al., 2017). However, with the prevalence of active learning systematic review tools (Harrison et al., 2020), active learning is at this point the most approachable machine learning method for systematic reviews, with the clearest benefits coming in the title and abstract screening phase (van de Schoot et al., 2021). By incorporating active learning, SYMBALS expedites the systematic review process while remaining accessible.

Table 2.1 provides an overview of the discussed papers that present a systematic review methodology. Methodologies that include automation techniques will generally be swifter, but accessibility can suffer. These methodologies can be less accessible due to their reliance on techniques and tooling that is not freely and publicly available, as is the case for the reinforcement learning approach of Ros et al. (2017). Additionally, since many researchers still do not fully trust automation techniques for systematic reviews (O'Connor et al., 2019), methodologies using these techniques are less accessible in the sense of being less approachable. One way to solve this issue is to incorporate trusted systematic reviews methods such as snowballing, as we propose to do with SYMBALS. Table 2.1 shows that a methodology that manages to be both accessible and swift is unique. Therefore, if SYMBALS manages to foster accessibility and swiftness, it has the potential to be of added value to the research community.

2.3 SYMBALS

In this chapter, we introduce SYMBALS: a SYstematic review Methodology Blending Active Learning and Snowballing. Figure 2.2 presents our methodology. Focusing on the planning and conducting phases of a systematic review (Kitchenham and Charters, 2007), SYMBALS complements existing review elements with active learning and snowballing steps. The following sections outline the steps that together constitute SYMBALS.

2.3.1 *Develop and evaluate protocol*

Any systematic review is instigated from a motivation and a need for the review (Wohlin, Runeson, et al., 2012). These lead to the formulation of research questions and the design of a systematic review protocol (Kitchenham and Charters, 2007). A protocol for SYMBALS should contain the following items:

- Background, rationale, and objectives of the systematic review.
- Research questions the systematic review aims to answer.
- Search strategy to be used.
- Selection criteria to be applied.
- Selection procedure to be followed.
- Data extraction, management, and synthesis strategy.
- Validation method(s) used to validate the procedure and the results.

Quality assessment checklists and procedures (Kitchenham and Charters, 2007) are vital to include if one plans to apply a quality assessment step. However, it is recognised that this is not a necessary phase in all systematic reviews (Brereton et al., 2007). Additional items that can potentially be included in a protocol are the risks of bias in the primary studies and the review itself (Moher et al., 2015), as well as a project timetable and dissemination strategy (Kitchenham and Charters, 2007; Wohlin, Runeson, et al., 2012).

For researchers in the field of information systems and other comparable fields, it is important to be aware of two potential roadblocks to implementing our methodology. Firstly, not all databases are designed to support systematic reviews (Brereton et al., 2007), meaning researchers may need to apply different search criteria in different sources. Secondly, abstracts in the information systems field are often of a quality that is too poor to be relied upon when applying selection criteria (Brereton et al., 2007). This problem can be circumvented by additionally inspecting the conclusions of these papers, and we have not found this issue to extensively impact the effectiveness of the active learning phase of SYMBALS.

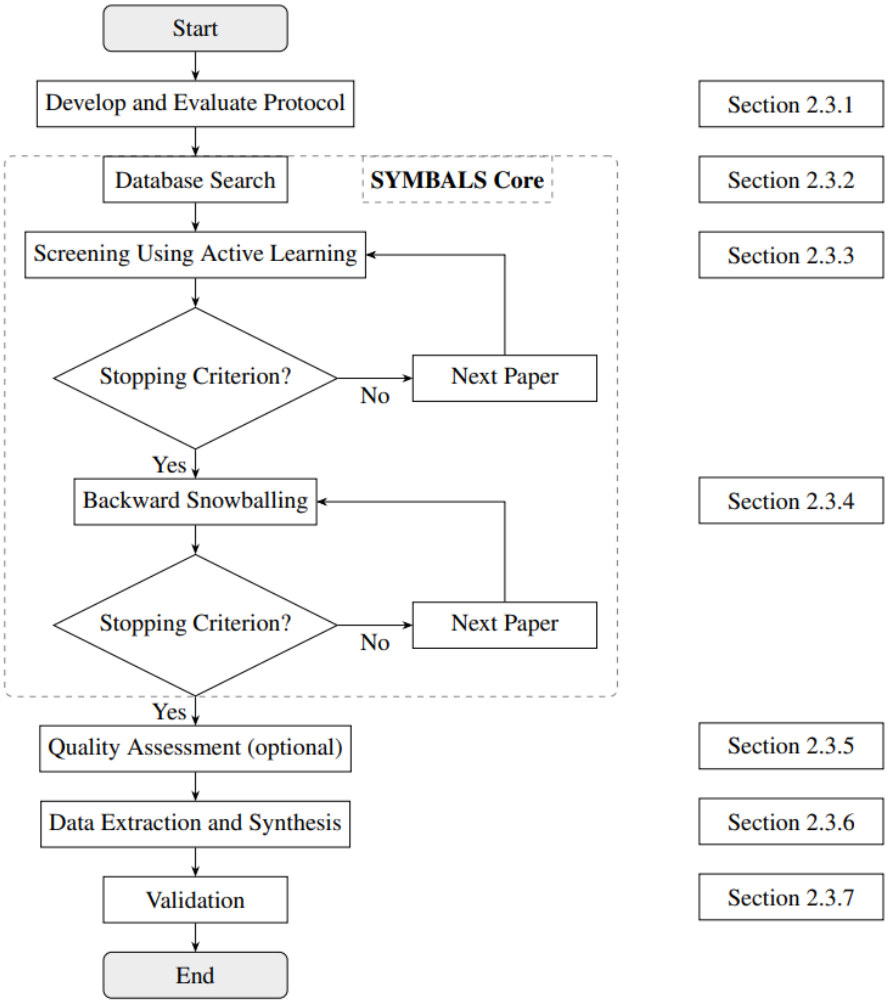


Figure 2.2: SYMBALS, our proposed systematic review methodology. The methodology consists of the SYMBALS core (dashed box), supplemented with elements of the stages of planning and conducting a review (Kitchenham and Charters, 2007).

2.3.2 *Database search*

Once researchers are content with their search string selection, they can start with the database search step of SYMBALS. Techniques exist to aid researchers in formulating their search query (Marcos-Pablos and García-Peñalvo, 2018), even involving machine learning methods (Marcos-Pablos and García-Peñalvo, 2020). We highly recommend researchers consult these methods to help in swiftly constructing a suitable search string.

The advantage of SYMBALS is that the search string does not need to be perfect. Not all databases offer the same search capabilities (P. Singh and K. Singh, 2017), meaning that complex, tailor-made search queries are often not reproducible across databases (Mourão, Kalinowski, et al., 2017). By using active learning, the impact of including papers that should not have been included is minimised. Concurrently, backward snowballing limits the impact of excluding papers that should have been included. By facilitating the use of a broad search query, SYMBALS is accessible for researchers without extensive experience in the field being considered. This is not only a benefit to junior researchers and students but also to researchers looking to map findings from other areas to their field of interest.

Different databases are relevant in different disciplines, and the set of relevant databases is bound to change over time. This is the reason that we do not recommend a fixed set of databases for our approach. Nevertheless, a few points are worth noting regarding the choice of database. Generally, there is a consensus of which databases are relevant to a particular field (Brereton et al., 2007; Kitchenham and Charters, 2007), and research has shown which databases are suitable for systematic reviews (Gusenbauer and Haddaway, 2020). Additionally, researchers should be aware of the required data of the active learning tool they intend to use for screening.

2.3.3 *Screening using active learning*

In the active learning phase, we recommend using existing and freely accessible active learning tools that are aimed at assisting title and abstract screening for systematic reviews. Researchers can consult tool evaluations (Harrison et al., 2020) to decide for themselves which tool they prefer to use. Although even the tools specifically aimed at automating systematic reviews suffer from a lack of trust by researchers (O'Connor et al., 2019), we believe that initiatives such as those to make code available open-source (van de Schoot et al., 2021) will solve many of the trust issues in the near future.

It is difficult to choose an appropriate active learning stopping criterion (Yu and Menzies, 2019). Some tools choose to stop automatically when the algorithm classifies none of the remaining papers as relevant (Wallace et al., 2012). Although this accommodates reproducibility, it is generally not acceptable for researchers to have no control over when they are done with

their screening process. Commonly used stopping criteria are to stop after evaluating n irrelevant papers in a row or after having evaluated a fixed number of papers (Ros et al., 2017). The simplicity of these stopping criteria is pleasant, but these criteria are currently not considered best practice (Yu and Menzies, 2019).

Of particular interest are those criteria that are based on an estimate of the total number of relevant papers in the starting set (Cormack and Grossman, 2016). Let N be the total number of papers and R the number of relevant papers. In general, R is not known. To estimate R we can evaluate papers until we have marked r papers as relevant. Let i denote the number of papers that are marked as irrelevant at this stage. We can then estimate R as:

$$R \approx N \times \frac{r}{r+i}. \tag{2.1}$$

A potential stopping criterion is then to stop once a predefined percentage p of the estimated number of relevant papers R has been marked relevant. This criterion solves the issues that the earlier criteria faced. Implementations of this approach that are more mathematically grounded exist (Cormack and Grossman, 2016; Yu and Menzies, 2019), and we encourage researchers to investigate those methods to decide on their preference.

2.3.4 *Backward snowballing*

There are systematic review methods that move straight to the quality assessment stage after applying active learning (Yu and Menzies, 2019). In SYMBALS we choose to blend active learning and backward snowballing. This allows researchers to complement their set of relevant papers with additional sources. There are three main classes of relevant papers that may not be included at this stage. The first is the group of relevant papers included in the set that was automatically excluded in the active learning phase. An appropriately defined stopping criterion should keep this set relatively small. Additionally, there are relevant papers that do not satisfy the search query used. Last, and certainly not least, is the group of relevant papers that are not present in the databases considered. This will mostly be grey literature and, from our experience, relatively old research.

Altogether these groups form the motivation to include a snowballing step, and it has been shown that this step has the potential to add many relevant papers, even after a database search (Mourão, Pimentel, et al., 2020). Additional relevant research can be identified from the reference lists (backward snowballing) and citations (forward snowballing) of included papers (Wohlin, 2014). After constructing an initial set of relevant inclusions and defining a stopping criterion, the backward snowballing procedure begins. In SYMBALS, the set of inclusions to consider is the set originating from the active learning process. This set will generally be much larger than the initiating set of a

regular snowballing procedure (Wohlin, 2014). This makes it vital to define a suitable stopping criterion, to prevent the backward snowballing process from taking up too much time.

Figure 2.3 depicts the backward snowballing procedure in our setting. The procedure differs from the traditional backward snowballing procedure (Wohlin, 2014) due to the large set of inclusions that already exist in our process from the active learning phase. This also implies the stopping criterion for backward snowballing has to differ from traditional stopping criteria (Wohlin, 2014). One could consider stopping after evaluating n irrelevant references or papers in a row. We recommend stopping when in the last N_r references, the number of new relevant additions r_r is less than some constant C, given that the number of snowballed papers s is at least S. For example, if our set of inclusions contains 100 papers, we may set the minimum number of papers to snowball to $S = 10$. Once 10 papers have been snowballed, we stop when the last $N_r = 100$ references contained less than $C = 5$ additions to our inclusions.

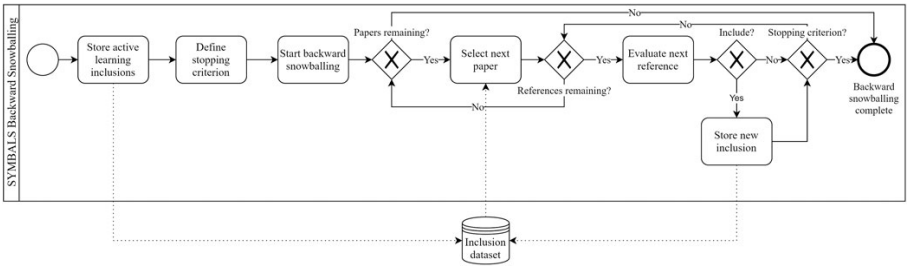


Figure 2.3: The backward snowballing process in the SYMBALS setting, depicted using BPMN. Although our process clearly differs from the traditional backward snowballing process, the diagram is undeniably similar to conventional snowballing diagrams (Wohlin, 2014).

Although both backward snowballing and forward snowballing can be potentially relevant, we argue to only apply backward snowballing in SYMBALS. Given that grey literature and older papers will generally constitute the largest group of relevant papers not yet included, it is more apt to inspect references than citations. Forward snowballing is well suited to updating systematic reviews (Wohlin, Mendes, et al., 2020), but, as we show in Section 2.4.1.7, SYMBALS can also be used to update a systematic review.

2.3.5 *Quality assessment*

From the core of SYMBALS, we now move back to traditional stages in systematic review methodologies. It is common to apply a quality assessment procedure to the research included after the completion of title and abstract

screening (Kitchenham and Charters, 2007). It is certainly not a mandatory step in a systematic review (Brereton et al., 2007), nor is it a mandatory step.

Based on criteria for good practice (Kitchenham, S. L. Pfleeger, et al., 2002), the software engineering field outlines four main aspects of quality assessment: reporting, rigour, credibility, and relevance (Dybå and Dingsøyr, 2008). We believe these aspects to be broadly applicable. According to the specific needs of a systematic review, quality criteria can be formulated based on the four main aspects (Y. Zhou et al., 2015).

No universally accepted quality assessment methodology exists (Zhang and Ali Babar, 2013). Automation of quality assessment is generally not even discussed. This highlights that there are possibilities to improve current quality assessment practice with machine learning techniques.

2.3.6 *Data extraction and synthesis*

Researchers should design data extraction and collection forms (Kitchenham and Charters, 2007) based on the research questions formulated during protocol development. These forms have the express purpose of helping to answer the research questions at hand but can also facilitate verifiability of the procedure. A well-designed data extraction form can even be made publicly available in conjunction with a publication (Morrison et al., 2018), to stimulate further research based on the results.

Data synthesis involves either qualitatively or quantitatively summarising the included primary studies (Kitchenham and Charters, 2007). Quantitative data synthesis, or meta-analysis, is especially useful if the extracted data is homogeneous across the included primary studies (Wohlin, Runeson, et al., 2012). Homogeneity can be promoted through a well-defined data extraction form. When performing a meta-analysis, researchers should be careful to evaluate and address the potential for bias in the primary studies (Wohlin, Runeson, et al., 2012), as this can threaten the validity of the results. It is recommended to include quality assessment results in the data synthesis phase, as it can offer additional insights into the results obtained by primary studies of varying quality.

2.3.7 *Validation*

The last step in our methodology is validation. Although validation is not explicitly included in all systematic review methodologies (Kitchenham and Charters, 2007; Wohlin, Runeson, et al., 2012), its importance is clearly recognised (Brereton et al., 2007; Moher et al., 2015). It is quite common for systematic reviews to assess the quality of primary studies based on whether limitations and threats to validity are adequately discussed (Y. Zhou et al., 2015). We want to promote validation in systematic reviews themselves, which

is why validation is a separate step in SYMBALS, rather than simply another reporting item.

There are four main validity categories: construct, internal, external, and conclusion (X. Zhou et al., 2016). We designed our methodology to counter threats to validity from all categories. Examples are unclear inclusion and exclusion criteria (Khan et al., 2019) and a subjective quality assessment (X. Zhou et al., 2016). Other commonly included elements during validation are an estimate of coverage of relevant research (Zhang, Babar, et al., 2011) and an investigation of bias handling in data extraction and synthesis (X. Zhou et al., 2016).

The swiftness of our methodology allows us to introduce a new validation method in this chapter: replication. An application of this novel validation method is presented in Section 2.4.1.7.

2.4 CASE STUDIES

To assess the properties and the validity of our methodology, we performed two case studies. The first investigates the ability of SYMBALS to accommodate both broad coverage and a swift process. The second compares our methodology to the FAST^2 (Yu and Menzies, 2019) methodology on four benchmark datasets. This allows us to evaluate both the effectiveness of our methodology in an absolute sense (case study 1) and relative to a state-of-the-art methodology (case study 2).

In both case studies, we used ASReview (van de Schoot et al., 2021) to perform title and abstract screening using active learning. Besides the fact that we found this tool to be easy to use, we applaud the commitment of the developers to open science and welcome their decision to make the codebase available open-source. Nonetheless, we want to stress that there are many other potent active learning tools available (Harrison et al., 2020).

As with most tools that support active learning for title and abstract screening, ASReview offers many options for the model to use (van de Schoot et al., 2021). We elected to use the default Naïve Bayes classifier, with TF-IDF feature extraction and certainty-based sampling. The authors state that these default settings produced consistently good results across many datasets (van de Schoot et al., 2021). Since Naïve Bayes is generally considered to be a relatively simple classifier, and the default feature extraction and sampling settings are available in most other active learning tools (van de Schoot et al., 2021), using these default settings facilitates reproducibility of our results.

2.4.1 *Case study 1: cybersecurity metric research*

The field of cybersecurity needs to deal with a constantly changing cyber threat landscape. Security practitioners and researchers feel the need to ad-

dress this challenge by devising security solutions that are by their nature adaptable (Sengupta et al., 2020; C. Wang and Lu, 2018). This requires a corresponding adaptivity in cybersecurity research methods, which is why cybersecurity metric research is an appropriate domain to apply and examine our approach.

Although research into the measurement of cybersecurity risk has matured in past decades, it remains an area of fierce debate. Some researchers feel that quantified security is a weak hypothesis, in the sense that "it lacks clear tests of its descriptive correctness" (Verendel, 2009). Others feel it is challenging, yet feasible (S. Pfleeger and Cunningham, 2010). Yet others conjecture that security risk analysis does not provide value through the measurement itself, but through the knowledge analysts gain by thinking about security (Slayton, 2015). Nevertheless, the overwhelming consensus is that cybersecurity assessment is necessary (Jaquith, 2007).

Reviews are common in the cybersecurity metric field, but they are generally not systematic reviews. There are exceptions, although most are either outdated at this stage (Rudolph and Schwarz, 2012; Verendel, 2009), or only cover a specific area of cybersecurity, such as incident management (Cadena et al., 2020). In a particularly positive exception in the area of software security metrics (Morrison et al., 2018), the researchers did not only provide a clear explanation of their methodology but have also made their results publicly available and accessible. Still, there is a need for a broad systematic review in this area, and with this first demonstration and future research, we hope to build on initial positive steps.

In the interest of brevity, we will only cover those facets and findings of our application that are of general interest, leaving out specific details of this implementation.

2.4.1.1 *Develop and evaluate protocol*

The first step in SYMBALS is to develop and evaluate a systematic review protocol. Our protocol was constructed by one researcher and evaluated by two others. Based on existing guidelines on relevant databases (Kitchenham and Charters, 2007), we selected the sources depicted in Figure 2.4. CiteSeerx and JSTOR were excluded due to the inability to retrieve large quantities of research from these sources. The search string selected for the Scopus database was:

```
AUTHKEY((security* OR cyber*)
AND (assess* OR evaluat* OR measur* OR metric* OR model* OR risk*
    OR scor*))
AND LANGUAGE(english) AND DOCTYPE(ar OR bk OR ch OR cp OR cr OR re
    )
```

The asterisks denote wildcards. We only considered English language publications and restricted the search to articles (ar), books (bk), book chapters (ch), conference papers (cp), conference reviews (cr) and reviews (re).

2.4.1.2 *Database search*

The Scopus search string did not always translate well to other databases. This is a known issue (P. Singh and K. Singh, 2017) which we cannot fully circumvent, although a simpler search string helps to solve this problem. Other problems we encountered were that ACM Digital Library and IEEE Xplore limit the number of papers you can reasonably access to 2,000 and that IEEE Xplore only allows the use of six wildcards in a query. In the end, we chose to stick with our original query and sources, knowing that the active learning and snowballing phases would help in solving most of the potential issues. After cleaning and deduplication, 25,773 papers remained.

Figure 2.4: The SYMBALS implementation for the cybersecurity metric research case study. The database search, screening using active learning, backward snowballing, and quality assessment steps are shown, with the number of inclusions at each stage.

2.4.1.3 *Screening using active learning*

For the active learning phase, we used ASReview (van de Schoot et al., 2021). We elected to stop evaluating when 20 consecutive papers were marked irrelevant; a simple criterion similar to criteria used in earlier work (Ros et al., 2017). Figure 2.4 shows that 1,644 papers remained at the end of the active learning phase.

2.4.1.4 *Backward snowballing*

Next, we applied backward snowballing. We copied the evaluation order of the active learning phase. This is a simple and reproducible strategy, that we recommend others to follow when applying our methodology. We chose

Table 2.2: The quality criteria applied to 60 papers during the quality assessment phase. The most commonly used criteria (Y. Zhou et al., 2015) were assessed for relevance. The most relevant criteria were reformulated to be suitable for use in combination with a Likert scale. Statements could be responded to with strongly disagree (SD), disagree (D), neutral (N), agree (A), or strongly agree (SA).

ASPECT	CRITERION	SD	D	N	A	SA
Reporting	There is a clear statement of the research aims.	0	4	7	28	21
	There is an adequate description of the research context.	0	6	11	17	26
	The paper is based on research.	0	3	3	16	38
Rigour	Metrics used in the study are clearly defined.	0	10	19	16	15
	Metrics are adequately measured and validated.	1	24	22	8	5
	The data analysis is sufficiently rigorous.	0	21	17	14	8
Credibility	Findings are clearly stated and related to research aims.	0	8	19	25	8
	Limitations and threats to validity are adequately discussed.	30	18	8	2	2
Relevance	The study is of value to research and/or practice.	0	9	12	28	11

to stop when 10 consecutive papers contained no additions to our set of inclusions; a strict but simple criterion. If researchers are looking for an alternative strategy, we recommend considering a stopping criterion based on the inclusion rate over the last N_r references, where N_r is a predefined constant. An example of such a strategy is given in Section 2.3.4. The backward snowballing phase left 1,796 included papers.

2.4.1.5 *Quality assessment*

Given the large number of included papers at this stage, the logical choice was to apply a quality assessment step. We adapted the most relevant commonly used quality criteria (Y. Zhou et al., 2015), to be suitable for use in combination with a Likert scale. Two researchers evaluated 40 papers each, with 20 of those papers being evaluated by both researchers. Table 2.2 shows the averaged results, where the scoring of the first researcher was used for the 20 duplicate papers.

The response to each quality criterion was scored with 0, 0.25, 0.5, 0.75 or 1, corresponding to the five possible evaluations. With the sheer size of the set of inclusions, it was not possible to assess the quality of all papers. One possible solution to this problem is the following. We split the 60 evaluated papers into a training set (48 papers) and a test set (12 papers). Each paper was labelled as having sufficient quality if it obtained a score of at least 6 out of 9. In the 20 papers that were evaluated by both researchers, there were 5 edge cases where a disagreement occurred. On average, the quality scores differed by roughly 0.7 points. The researchers were almost equally strict in the evaluation of the papers, with the total sum of all quality scores differing by just 0.25.

We extended our quality scores with three explanatory features: years since publication, citation count, and the number of pages. A binary decision tree was trained on the explanatory features for the 48 training papers and evaluated on the 12 test papers. The model predicted 11 of the 12 papers correctly, incorrectly predicting one edge case with a quality score of 6 as having insufficient quality.

This short demonstration shows that training decision trees on assessed papers is a viable alternative to other strategies to filter a large set of inclusions. Commonly used alternatives are to only consider articles or to limit the time frame of the search. A decision tree trained on actual researcher quality assessments is an interesting substitute for traditional approaches, although we wish to stress that it is fully up to researchers using SYMBALS to choose which approach they apply. Additionally, quality assessment is an optional phase in SYMBALS, meaning researchers could even choose to not apply this step.

2.4.1.6 *Data extraction and synthesis*

After applying the resulting criteria of the decision tree to our inclusions, the 516 inclusions indicated in Figure 2.4 remained. The set of excluded papers comprised both research that did not pass the decision tree assessment and research that had insufficient data for assessment. Figure 2.5 illustrates the importance of the backward snowballing phase. Of our inclusions, 17% originated from backward snowballing. Considering only papers from before 2011, this figure jumps to 45%, highlighting the potential weakness of using only a database search step. Figure 2.5 therefore demonstrates the ability of SYMBALS to ensure broad coverage over time.

After an initial analysis of our inclusions, we formulated our data extraction form and used this as a guide to extract the necessary data. We then used quantitative data synthesis to produce more detailed and insightful results, aided by the homogeneity of our extracted data. Given that this is a demonstration of our methodology, rather than a complete systematic review study, we leave further analysis and presentation of our detailed results for future work.

2.4.1.7 *Validation*

To validate our case study, as well as the methodology itself, we performed a replication experiment. We extended the existing review with research from the months following the initial database search, using the same initiation process and stopping criteria as defined in Sections 2.4.1.3 and 2.4.1.4. The replication was performed by both the main researcher and a researcher who was not involved in the initial review. This allowed us to answer the question of whether SYMBALS contributes to an accessible and swift process.

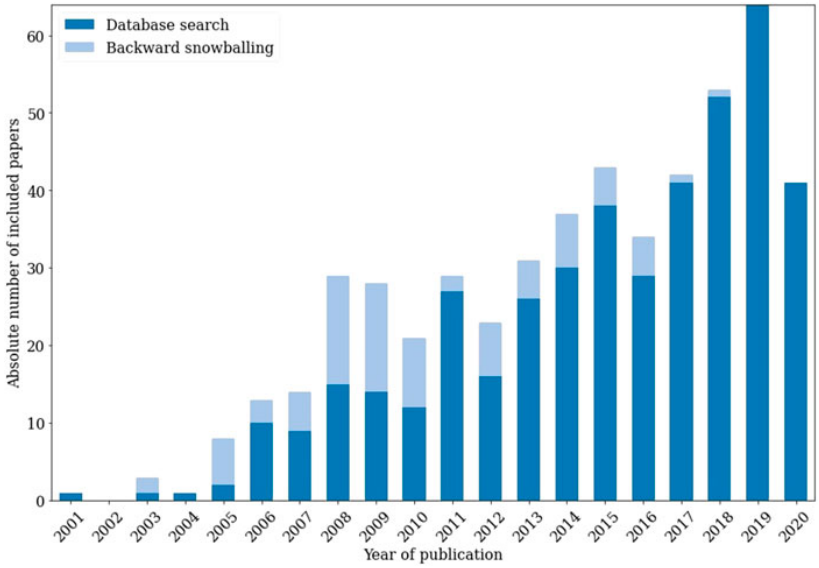


Figure 2.5: The absolute number of cybersecurity metric papers per year in the final inclusion set. We distinguish papers resulting from database search (dark) from those resulting from backward snowballing (light). For papers from 2010 and earlier, 45% originated from backward snowballing.

The database search procedure uncovered 2,708 papers, of which 222 were evaluated in the active learning phase. In the backward snowballing phase the main researcher evaluated 300 references. A common estimate for the time taken to screen a title-abstract record is a minute (Shemilt et al., 2016). This aligns with our time spent on the screening phase, which was 4 hours (222 minutes is 3.7 hours). The average time to scan one reference during backward snowballing can be expected to be lower than a minute, since a certain portion of the references will either have been evaluated already or will be obviously irrelevant (e.g., website links). Our backward snowballing phase took 3.5 hours, which corresponds to 0.7 minutes per reference. Altogether the process took 7.5 hours, whereas screening the titles and abstracts of 2,708 papers would have taken over 45 hours. Hence, we were able to speed up the title and abstract screening phase by a factor of 6.

To address the question of accessibility, we asked a researcher that had not been involved in the review to also perform the replication experiment. After 2 hours of explanation, the researcher was able to complete the active learning and snowballing phases, albeit roughly 3 times as slow as the main researcher. Note that this is still twice as fast as the traditional process. Automatic exclusion during active learning contributes to this speed. However, given the relatively short time that was required to explain the methodology, we argue

that the structure SYMBALS offers is another reason that it accommodates a swift process.

An additional element that is worth addressing is trust in the active learning process (O'Connor et al., 2019). One question that hovers over machine learning techniques is whether their random elements negatively impact reproducibility. To test this statement for the ASReview tool, we investigated how the first 100 papers of the active learning phase would change under different levels of disagreement with the main researcher. Our ASReview process starts after presenting 5 prior relevant papers to the tool and evaluating 5 random papers. In our first experiment, we copied all earlier decisions by the main researcher. This already resulted in small changes to the order in which papers were recommended. This poses a problem when using our stopping criterion, as changes in the order can alter the moment at which a researcher has reached n consecutive irrelevant papers. This is one of the reasons we recommend using more sophisticated stopping criteria.

The changes in order persisted when for 20% of the papers the initial evaluation of the main researcher was reversed. In both cases, the changes in order were minimal for the first 20 papers. This is important, as these papers will be the first papers considered in the backward snowballing phase. The replication of the second researcher had an even higher level of disagreement in the first 100 papers of 37%, which was a natural consequence of differing experience in the cybersecurity metrics field. Interestingly, even with this level of disagreement, the first 17 papers did not contain a paper outside of the first 25 papers of the main researcher. We believe this shows that the process is robust to inter-rater disagreement, given the correct stopping criterion.

2.4.2 *Case study 2: benchmarking*

Besides evaluating the performance of our methodology in an absolute sense, we additionally evaluated its performance compared to an existing state-of-the-art methodology. We benchmarked the SYMBALS methodology using datasets (Yu, Barik, et al., 2020) developed for the evaluation of the FASTREAD (Yu, Kraft, et al., 2018) and FAST^2 (Yu and Menzies, 2019) systematic review methodologies. The datasets of both inclusions and exclusions were constructed based on three systematic reviews (Hall et al., 2012; Radjenović et al., 2013; Wahono, 2007) and one tertiary study (Kitchenham, Pretorius, et al., 2010).

In our benchmarking, we compare to the results obtained by the FAST^2 methodology, since it is an improvement over the FASTREAD methodology (Yu and Menzies, 2019). For the three systematic reviews (Hall et al., 2012; Radjenović et al., 2013; Wahono, 2007), the authors reconstructed the datasets based on information from the original papers. For the tertiary study (Kitchenham, Pretorius, et al., 2010), the dataset was provided by the original authors of the review. The reason that we chose to compare to FAST^2 is not

only because it is a state-of-the-art methodology, but also because the FAST^2 datasets were so easily accessible and in a compatible format for SYMBALS. This was not the case for the other methodologies covered in Table 2.1, such as Mourão, Kalinowski, et al. (2017) and Mourão, Pimentel, et al. (2020).

SYMBALS and FAST^2 cannot be fairly compared without first adjusting the datasets. After a database search, the FAST^2 method uses active learning as the sole approach for title and abstract screening. In the FASTREAD and FAST^2 papers, the authors make the necessary assumption that the datasets encompass all relevant papers since these methodologies have no way of discovering relevant research outside of the original dataset. However, in research that incorporates snowballing in systematic reviews, it has been shown that between 15% and 30% of all relevant papers are not included in the original dataset (Mourão, Kalinowski, et al., 2017; Mourão, Pimentel, et al., 2020). This aligns with our results in the first case study, where 17% of the inclusions originated from backward snowballing.

To enable a fair comparison of SYMBALS and FAST^2, we randomly removed 15% of both the relevant and irrelevant papers in the datasets before initiating our active learning phase. The removed papers were then considered again in the backward snowballing phase of SYMBALS. This adjustment allows our benchmarking study to accurately reflect the actual situation faced by researchers performing systematic reviews. The consequence of this adaptation is that the recall achieved by the FAST^2 methodology is multiplied by a factor of 0.85.

Both the FASTREAD and FAST^2 papers address the definition of an initiation process and a stopping criterion. Regarding initiation, two approaches are posited: 'patient' and 'hasty.' The patient approach generates random papers and initiates active learning once 5 inclusions are found. The hasty approach initiates active learning after just 1 inclusion is found. To leave room for the backward snowballing phase, we used the hasty method for initiation.

Many of the stopping criteria considered in FAST^2 cannot be applied in our setting, since they rely on properties of the specific active learning tool used for the methodology. To ensure a transparent approach, we opted to stop after 50 consecutive exclusions. This stopping criterion, sourced from earlier work (Ros et al., 2017), was found to yield the fastest active learning phase on average in the FAST^2 paper. This is useful in our setting, as it again leaves time for the backward snowballing phase.

We conducted the active learning phase of our benchmarking experiments using the ASReview tool (van de Schoot et al., 2021) that we also used in our first case study. The results are shown in Figure 2.6. As mentioned before, the recall achieved by the FAST^2 methodology was multiplied by a factor of 0.85, to align with the removal of 15% of the papers.

The FAST^2 results are linear interpolations of the median results provided by the authors in their paper. For the later data points, this linear extrapolation represents the actual data with reasonable accuracy. However, for the

earlier data points, the linear extrapolation overestimates the recall achieved by FAST^2. FAST^2, like SYMBALS, takes time to find the first few relevant papers, due to the nature of the applied initiation process. This observation is confirmed when examining the graphs presented in the FAST^2 paper. Although the overestimation of recall in the early phase is not ideal for our comparison, we are mainly interested in how the methods compare beyond initiation. We employ the same initiation process as FAST^2, meaning differences in performance during the initiation phase are purely due to random deviations.

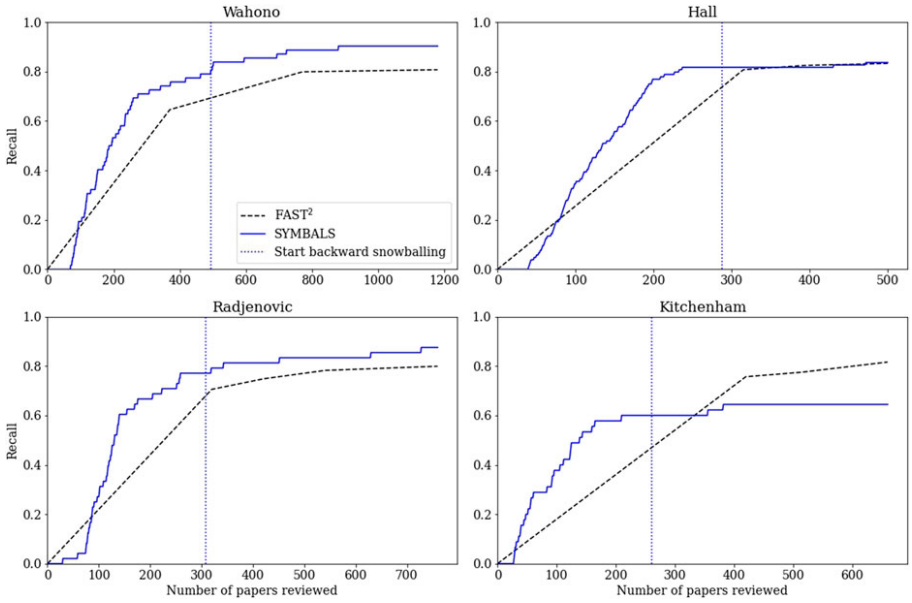


Figure 2.6: The recall achieved by the FAST^2 (Yu and Menzies, 2019) and SYMBALS methodologies, for the four review datasets studied in our benchmarking case study. For the FAST^2 method we provide linear interpolations of the median results. A vertical dotted line indicates the start of the backward snowballing phase for SYMBALS.

For the three traditional systematic review papers (Hall et al., 2012; Radjenović et al., 2013; Wahono, 2007), our methodology achieved a higher recall than FAST^2. At the maximum number of reviewed papers, SYMBALS achieved a 9.6% higher recall for the Wahono dataset (90.3% compared to 80.7%), a 0.4% higher recall for the Hall dataset (83.7% compared to 83.3%), and a 7.6% higher recall for the Radjenovic dataset (87.5% compared to 79.9%). In all three of these cases, the active learning phase of SYMBALS performed well, achieving a recall higher than the recall of FAST^2 after evaluating the same number of papers. Nevertheless, in each case, the recall achieved after the active learning phase was lower than the eventual recall of FAST^2.

The backward snowballing phase of our methodology raised the recall achieved in the active learning phase by 9.7% for the Wahono dataset, by 1.9% for the Hall dataset, and by 10.4% for the Radjenovic dataset. At first, these contributions may seem to be minor. However, as recall increases, further improving recall becomes increasingly difficult. In light of this observation, the backward snowballing additions are the key element in ensuring that SYMBALS outperforms FAST^2 for the Wahono, Hall, and Radjenovic datasets. Considering the finding from our first case study that reviewing references during backward snowballing is faster than screening titles and abstracts during active learning, SYMBALS achieves a higher recall in less time than FAST^2.

For the tertiary study (Kitchenham, Pretorius, et al., 2010), the performance of SYMBALS (64% recall) was relatively poor compared to FAST^2 (82% recall). Both the active learning phase and the backward snowballing phase underperformed compared to the other studies. Regarding the active learning phase, one explanation could be that the content of the titles and abstracts were not identifiably different for relevant and irrelevant papers. This is certainly a plausible scenario given that the tertiary study screens systematic reviews, which are likely to differ more in their content than regular papers aimed at a specific topic. This does not explain, however, how FAST^2 was able to achieve a high recall. The difference between the performance of ASReview and the active learning of FAST^2 is a consequence of algorithmic differences, but these algorithmic differences were not investigated further.

It is not surprising that backward snowballing is less useful for tertiary studies, as the systematic reviews that they investigate are less likely to reference each other. Furthermore, systematic reviews often have many references. The 400 references we evaluated for the tertiary study, came from just 5 papers. With fewer papers to investigate, the scope of the backward snowballing phase is narrowed. A final factor that may have influenced results, is that the authors of the tertiary study explicitly focus on the period between the 1st of January 2004 and the 30th of June 2008. A short timespan restricts the effectiveness of backward snowballing.

We believe this benchmarking study highlights the areas where our approach can improve upon existing methodologies. When researchers are looking to systematically review research over a long period, SYMBALS can trump state-of-the-art methodologies on their home turf. When researchers are interested in additionally including grey literature or expect that not all relevant papers are included in their initial dataset, our methodology offers further advantages through the inclusion of a backward snowballing step. When researchers are performing a tertiary study, fully automated methods such as FAST^2 may be more appropriate than SYMBALS. Future research employing and evaluating our methodology will help to further clarify its strengths and weaknesses.

2.5 DISCUSSION AND LIMITATIONS

We posed the following research question at the outset of this chapter: How can active learning and snowballing be combined to create an accessible and swift systematic review methodology? The review of existing research in systematic review methodologies and active learning in Section 2.2, combined with the additional analysis in Sections 2.3.3 and 2.4.1.4, helped us to formulate a methodology inspired and motivated by existing work. Figure 2.2 outlines the resulting proposal. We found that active learning is best suited to the screening of titles and abstracts and that backward snowballing provides an ideal supplement. The combination facilitates coverage of relevant (grey) literature while maintaining a reproducible procedure.

In the case study of Section 2.4.1, 17% of the relevant research would not have been found without backward snowballing. This figure jumps to 45% when only considering research from before 2011. We further investigated the properties of our methodology in Section 2.4.1.7. The fact that a researcher who was new to the case study review was able to execute our methodology after just two hours of explanation, shows that it is easily understandable and accessible. Moreover, SYMBALS was shown to accelerate title and abstract screening by a factor of 6, proving that it accommodates a swift procedure through its active learning component.

Section 2.4.2 compared the performance of our approach to the state-of-the-art systematic review methodology FAST^2 (Yu and Menzies, 2019). We found that SYMBALS achieves a 6% higher recall than FAST^2 on average when applying the methodologies to systematic reviews. FAST^2 was found to outperform SYMBALS for a tertiary study benchmark, pointing to a possible case where SYMBALS may not be the most suitable methodology.

Our methodology has its limitations. The lack of trust in systematic review automation technologies (O'Connor et al., 2019) is not fully solved by SYMBALS. Active learning methods and tools have matured, but there will still be researchers who feel uncomfortable when applying them in reviews. This limits the use of our approach to only those researchers who trust the automation technologies employed. Likewise, practical limitations exist. Depending on the exact implementation, researchers will have to have some computer programming skills. ASReview, for example, requires the installation and use of the ASReview Python package. The heterogeneity of online databases is another limitation our methodology cannot fully address, although the fact that SYMBALS allows researchers to avoid complex search queries partially counters this issue.

Lastly, we should address potential threats to validity. A handful of researchers evaluated SYMBALS throughout this process. Although their varying experience levels and areas of expertise allowed us to address questions of accessibility and reproducibility, we admit that in the future more evaluation is desirable. Another potential pitfall is the quality of abstracts in

fields outside the fields considered in our case studies. There are areas of research where it is known that abstract quality can be poor (Brereton et al., 2007). This can potentially harm the effectiveness of active learning in abstract screening. Altogether, we believe that the benefits of SYMBALS far outweigh its limitations, which is why we strongly believe it can have a lasting impact on the systematic review landscape.

2.6 CONCLUSION AND FUTURE RESEARCH

This chapter introduced SYMBALS: a SYstematic review Methodology Blending Active Learning and Snowballing. Our methodology blends the proven techniques of active learning and backward snowballing to create an effective systematic review methodology. A first case study demonstrated the ability of SYMBALS to expedite the systematic review process, while at the same time making systematic reviews accessible. We showed that our approach allows researchers to accelerate title and abstract screening by a factor of 6. The need for backward snowballing was established through its contribution of 45% to all inclusions from before 2011. In our benchmarking study we demonstrated the ability of SYMBALS to outperform state-of-the-art systematic review methodologies, both in speed and accuracy.

In future research, we hope to further evaluate and validate our methodology, including the completion of the full cybersecurity metric review case study. Another interesting avenue for future research is to investigate which choices in the selection of active learning tools, classification models, and stopping criteria are optimal in which scenarios. Optimising SYMBALS in these areas can certainly benefit researchers performing systematic reviews, although they should take care to not reduce the reproducibility of their results.

Finally, we believe that there are promising possibilities for further systematic review automation. Machine learning techniques and opportunities exist for all areas of the systematic review procedure. As these techniques mature, we will see an increase in their use. Research into how to incorporate these techniques in systematic review methodologies in a way that harbours trust, robustness, and reproducibility, is of paramount importance. We hope that SYMBALS is the next step in the right direction.

3

RESPITE FOR SMES: A SYSTEMATIC REVIEW

The contents of this chapter are based on: van Haastrecht, Yigit Ozkan, et al. (2021). Respite for SMEs: A systematic review of socio-technical cybersecurity metrics. Applied Sciences. For the full version including appendices, please consult the original article.

Cybersecurity threats are on the rise, and small- and medium-sized enterprises (SMEs) struggle to cope with these developments. To combat threats, SMEs must first be willing and able to assess their cybersecurity posture. Cybersecurity risk assessment, generally performed with the help of metrics, provides the basis for an adequate defence. Significant challenges remain, however, especially in the complex socio-technical setting of SMEs. Seemingly basic questions, such as how to aggregate metrics and ensure solution adaptability, are still open to debate. Aggregation and adaptability are vital topics to SMEs, as they require the assimilation of metrics into actionable advice adapted to their situation and needs. To address these issues, we systematically review socio-technical cybersecurity metric research in this chapter. We analyse aggregation and adaptability considerations and investigate how current findings apply to the SME situation. To ensure that we provide valuable insights to researchers and practitioners, we integrate our results in a novel socio-technical cybersecurity framework geared towards the needs of SMEs. Our framework allowed us to determine a glaring need for intuitive, threat-based cybersecurity risk assessment approaches for the least digitally mature SMEs. In future, we hope our framework will help to offer SMEs some deserved respite by guiding the design of suitable cybersecurity assessment solutions.

3.1 INTRODUCTION

In recent times, we have seen a surge in cyber threats that businesses are struggling to cope with (Bassett et al., 2021). Additionally, the frequency with which cybersecurity incidents occur, and the costs associated with them, are on the rise (Bissell and Lasalle, 2019). Among businesses, small- and medium-sized enterprises (SMEs) are most vulnerable, due to a shortage of cybersecurity knowledge and resources (Heidt et al., 2019). The vulnerable position of SMEs is being exploited, as witnessed by the large proportion of SMEs that experience cyber incidents (Ponemon Institute, 2019).

In SME cybersecurity, the interplay between the social and the technical is essential (Malatji, Von Solms, et al., 2019), which is why SMEs are often studied from a socio-technical systems (STS) perspective (Carías, Arrizabalaga, et al., 2020). The view of STS is that joint consideration of social and technical elements is necessary (Davis et al., 2014). This view has interesting implications in cybersecurity, where humans are generally found to be the weakest link (Gratian et al., 2018; Shojaifar, Fricker, and Gwerder, 2020).

Due to their lack of resources (Heidt et al., 2019) and the complex socio-technical setting they operate in, SMEs struggle to address their cybersecurity issues autonomously (Benz and Chatterjee, 2020). Before SMEs can begin to improve their cybersecurity posture, it is vital they first assess their current situation (Jaquith, 2007). Assessment of cybersecurity posture is achieved by measuring SME cybersecurity properties, which result in cybersecurity metrics. Regardless of whether measurement results are deemed relevant by the SME, the knowledge gained by those involved in the measurement process is of value (Slayton, 2015). This observation touches once more on the socio-technical nature of the problem, where furthering human knowledge and improving the technical cybersecurity posture of an SME go hand-in-hand.

Cybersecurity assessment generally requires the aid of cybersecurity experts; personnel that SMEs typically do not have (Benz and Chatterjee, 2020; Shojaifar, Fricker, and Gwerder, 2020). A solution to this issue is to automate the cybersecurity assessment process where possible (Shojaifar, Fricker, and Gwerder, 2020). Although automation is a promising approach, the diverse nature of the SME landscape is often ignored (European DIGITAL SME Alliance, 2020; Yigit Ozkan, Spruit, et al., 2019), whereas we know from earlier research that it is vital for SMEs to have solutions adapted to their context and needs (Cholez and Girard, 2014; Mijnhardt et al., 2016).

Another issue is that cybersecurity assessment approaches aimed at SMEs are still scarce (Carías, Arrizabalaga, et al., 2020), explaining why it is not uncommon to see results from other cybersecurity focus areas being applied to the SME setting (Benz and Chatterjee, 2020). Systematic literature reviews are a logical approach to gather knowledge from one focus area, summarise it, and make it available for use in other focus areas.

Systematic reviews that address both the social and technical sides of cybersecurity, already exist (J.-H. Cho et al., 2019; Pendleton et al., 2016). These reviews identified a need for adaptable solutions (J.-H. Cho et al., 2019), which we have seen are also craved by SMEs. Additionally, these papers stress the need for more clarity on how to aggregate security metrics (J.-H. Cho et al., 2019; Pendleton et al., 2016). Given the lack of resources available at SMEs, aggregating information into understandable insights is a requirement for a usable solution (Shojaifar, Fricker, and Gwerder, 2020).

The issue with these systematic reviews is that they offer adaptability and aggregation as areas for future research, rather than addressing the topics head-on. Additionally, they do not provide actionable insights for SMEs since this is not their target audience.

In short, we can conclude that SMEs need (semi-)automated cybersecurity assessment approaches that address their needs for adaptability and aggregation of information. A systematic review offers the potential to gather and summarise such information, providing guidelines for designing usable solutions for SMEs. This motivates the need for a systematic review of cybersecurity metric research, where both the social and technical sides of the puzzle are acknowledged. This is exactly our aim in this chapter, as we try to answer the following research questions:

- **RQ1**: How are cybersecurity metrics aggregated in socio-technical cybersecurity measurement solutions?
- **RQ2**: How do aggregation strategies differ in cybersecurity measurement solutions relevant to SMEs and all other solutions?
  - **RQ2.1**: What are the reasons for these differences?
  - **RQ2.2**: Which aggregation strategies can be used in SME cybersecurity measurement solutions, but currently are not?
- **RQ3**: How do cybersecurity measurement solutions deal with the need for adaptability?

In Section 3.2, we cover related work from several different perspectives to provide a basis for our systematic review. Our systematic review methodology is detailed in Section 3.3, after which we present our results in Section 3.4.

To ensure that the insights we gain on aggregation and adaptability are captured in an actionable form, we incorporate them in a novel socio-technical cybersecurity framework geared towards SME needs. Our framework, introduced in Section 3.5, integrates our systematic review results with existing knowledge to arrive at concise guidelines for what can be expected of various SME categories.

Section 3.6 focuses on outlining the answers to our research questions, as well as covering limitations and threats to validity. Finally, we conclude in Section 3.7, additionally outlining potentially fruitful areas for future research.

3.2 RELATED WORK

Before covering work relating to our socio-technical cybersecurity metric setting, we should be clear on our definition of what constitutes a cybersecurity metric. We make use of the definition of a cyber-system as specified in Refsdal et al. (2015): "A cyber-system is a system that makes use of a cyberspace." Refsdal et al. (2015) define cyberspace as "a collection of interconnected computerized networks, including services, computer systems, embedded processors, and controllers, as well as information in storage or transit." There is no standard definition of what constitutes a (cyber)security metric (Pendleton et al., 2016). Borrowing ingredients from earlier definitions, we define a cybersecurity metric to be *any value resulting from the measurement of security-related properties of a cyber-system* (Böhme and Freiling, 2008; Pendleton et al., 2016; Refsdal et al., 2015).

3.2.1 *Socio-technical cybersecurity*

Humans are often considered the weakest link in cybersecurity (Martens et al., 2019). It is vital to recognise the interaction of the social and technical sides of cyber-systems when modelling and measuring cybersecurity, which is why the field of STS has played such an important role in cybersecurity metric research (Gollmann et al., 2015). STS research has uncovered the dangers of considering social and technical elements separately (Selbst et al., 2019) and has offered insight into how to avoid these dangers (Davis et al., 2014).

Recognition of the human factor in cybersecurity goes beyond simply including static human actors. This is where behavioural theories such as Protection Motivation Theory (PMT) and Self-Determination Theory (SDT) come in (Menard et al., 2017; Padayachee, 2012). PMT reserves a prominent role for extrinsic motivators and threat appraisal (Herath and Rao, 2009). SDT includes extrinsic motivation as a central concept but often focuses on moving from extrinsic to increasingly internalised motivation (Padayachee, 2012). In the context of SMEs, intrinsic motivation to improve cybersecurity is often hard to find. However, there are solutions to this problem. Committing to improving cybersecurity in an organisation can motivate employees (Padayachee, 2012). From the STS perspective, it is common to distinguish between metrics that include the real-life threat environment and those that do not (Gollmann et al., 2015). Threat perception lies at the core of PMT and is important in security applications using SDT (Menard et al., 2017). Another solution to promote motivation among SME employees would therefore be to incorporate the real-life threat environment in our cybersecurity metrics. Later in this chapter, in Section 3.4, we describe whether this is indeed something we observe in current research.

We will address the social dimension using the ADKAR model of Hiatt (2006). This model, originating from change management, considers five

Table 3.1: Existing cybersecurity metric (systematic) reviews. The research focus area is shown, with 'generic' indicating research without a specific focus area. We consider social factors to be evaluated when the review covers socio-technical cybersecurity metrics.

RESEARCH	YEAR	FOCUS AREA	SOCIAL FACTORS
Current chapter	**2021**	**Generic**	✓
Verendel (2009)	2009	Generic	×
Rudolph and Schwarz (2012)	2012	Generic	×
Pendleton et al. (2016)	2016	Generic	✓
J.-H. Cho et al. (2019)	2019	Generic	✓
Husák, Komárková, et al. (2019)	2019	Attack Prediction	✓
Iannacone and Bridges (2020)	2020	Cyber Defense	×
Kordy et al. (2014)	2014	Directed Acyclic Graphs	×
Cadena et al. (2020)	2020	Incident Management	✓
Knowles et al. (2015)	2015	Industrial Control Systems	✓
Asghar et al. (2019)	2019	Industrial Control Systems	✓
Eckhart et al. (2019)	2019	Industrial Control Systems	×
Jing et al. (2019)	2019	Internet Security	×
Sengupta et al. (2020)	2020	Moving Target Defense	×
Liang and Xiao (2013)	2013	Network Security	×
Ramos et al. (2017)	2017	Network Security	✓
Cherdantseva et al. (2016)	2016	SCADA Systems	✓
Morrison et al. (2018)	2018	Software Security	×
W. He et al. (2019)	2019	Unknown Vulnerabilities	×
Xie et al. (2019)	2019	Wireless Networks	×

phases in managing the personal side of change: awareness, desire, knowledge, ability, and reinforcement. ADKAR has previously been applied in assessing information security culture within organisations (Da Veiga, 2018). We apply ADKAR as a means to classify the socio-technical cybersecurity metrics we encounter. We define a socio-technical cybersecurity metric to be *a cybersecurity metric that requires measuring the outcome(s) of the actions of at least one (simulated) human actor.* We do not address the technical dimension explicitly in this definition, as the technical dimension is implicit in the term 'cybersecurity.' We hypothesise that all socio-technical cybersecurity metrics can be linked to one or more of the ADKAR categories.

3.2.2 *Cybersecurity metric reviews*

Systematic reviews are common in cybersecurity metric research. However, as Table 3.1 shows, they are often narrow in scope. Either the focus area is narrow, or the research does not consider social factors. The papers that do cover both social and technical factors, often do so passingly, and without covering the intricacies and implications of socio-technical interactions.

Some exceptions are comprehensive and cover both social and technical factors (J.-H. Cho et al., 2019; Pendleton et al., 2016). Interestingly, exactly

these papers outline that future research should focus on "how to aggregate and to what extent to aggregate" (Pendleton et al., 2016). Additionally, they stress the importance of adaptability, meaning by this "the state of being able to change to work or fit better" (J.-H. Cho et al., 2019). This need for adaptability has been confirmed by experience from practice (Ray et al., 2020).

We address the acknowledged challenges of aggregation and adaptability head-on in our systematic review, ensuring that our approach is both distinct from earlier work and provides a meaningful contribution to the field. Furthermore, we employ a novel systematic review approach (as outlined in Section 3.3) and target our analysis to aid SMEs, a group with specific needs often not considered in earlier work.

3.2.3 *Aggregation*

In cybersecurity metric research, aggregation strategies vary, although the importance of proper aggregation is widely recognised (J.-H. Cho et al., 2019; Pendleton et al., 2016). To discuss different aggregation strategies, we define a mathematical context with an aggregation strategy $S : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$, where $\mathbb{R}_{\geq 0}$ is the set of non-negative real numbers. We define metric value variables x_i, corresponding to metrics $i = 1, \ldots, n$. The metric values are assumed to be non-negative: $x_i \in \mathbb{R}_{\geq 0}\ \forall i$. We assume that for each metric, a higher metric value corresponds to lower security, without loss of generality. A negative relationship between a metric and security is common in the security literature, as it is often the lack of security, or risk, which is being measured.

A desirable property of a strategy S is that it is responsive to changes in metric values. This is captured by the property of injectivity, where we consider a strategy S to be injective when for $a, b \in \mathbb{R}_{\geq 0}$, $a \neq b$, $S(a, x_1, x_2, \ldots, x_n) \neq S(b, x_1, x_2, \ldots, x_n)$. Injectivity implies that a change in a metric value will always result in a change of the aggregate, provided all else remains constant. A stronger requirement would be strict monotonicity of the strategy S. Although this property could be desirable in the cybersecurity context, we only consider the less strict injectivity in this chapter.

A common property of averages, which constitute a specific branch of aggregation, is idempotence. A strategy S is idempotent, when for $a \in \mathbb{R}_{\geq 0}$, $S(a, a, \ldots, a) = a$. When an aggregation strategy S is both injective and idempotent, the result of the aggregation always lies between the minimum and the maximum values of all metrics. Both injectivity and idempotence capture what we would intuitively expect of an aggregation strategy, as these are properties satisfied by the Pythagorean means. In this sense, these are desirable properties in the context of SMEs, where cybersecurity knowledge is often lacking. To still allow employees to feel competence and relatedness (Menard et al., 2017) in the complex cybersecurity setting, we should at least use an aggregation strategy they understand.

Three additional properties are important in the security context. The possibility to prioritise certain metrics over others is desirable (Lippmann and Riordan, 2016). Formally, we consider a strategy to allow for prioritisation when for any $a, b > 0$, $a \neq b$, there exists a pair i, j with $i \neq j$, such that $S(x_1, \ldots, x_i = a, \ldots, x_j = b, \ldots, x_n) \neq S(x_1, \ldots, x_i = b, \ldots, x_j = a, \ldots, x_n)$.

Strategies should also be able to accommodate dependencies between security metrics. However, it is complicated to include metric dependencies, with some seeing it as "the most challenging task" in aggregation (J.-H. Cho et al., 2019). For strategies in the set $\mathbb{D}$ of strategies that satisfy the necessary differentiability properties, we define a strategy S to allow for dependencies, when there exist distinct metrics i, j, and k such that:

$$\frac{\partial^2 S}{\partial x_i \partial x_j} \not\propto \frac{\partial^2 S}{\partial x_i \partial x_k}. \tag{3.1}$$

Equation 3.1 captures the idea that a strategy S allows for dependencies among metrics when it allows for relationships among metrics that are not proportional to other relationships. For aggregation strategies $S \notin \mathbb{D}$, we employ the same verbal definition. Care should be taken to adjust the criterion of Equation 3.1 appropriately where it cannot be applied directly for the strategy S.

A last core principle in security is that systems are only as secure as their weakest link (N. Ferguson and Schneier, 2003). Assuming that we have at least two distinct values among our metrics, there exists a minimum value x_{min} and a maximum value x_{max}. Since we assume metrics relate negatively to security, x_{max} corresponds to the weakest link. A strategy S satisfies the weakest link principle if for any $a > 0$, $S(x_{min} + a, \ldots, x_{max}) \leq S(x_{min}, \ldots, x_{max} + a)$, and there exists an $\alpha > 0$, such that $S(x_{min} + \alpha, \ldots, x_{max}) < S(x_{min}, \ldots, x_{max} + \alpha)$. Thus, weakening the weakest link has more impact than weakening the strongest link with an equal amount.

The most common aggregation strategy employed in the literature is the weighted linear combination (WLC), which can be defined as:

$$S_{WLC}(\mathbf{x}) = a + \frac{\sum_{i=1}^{n} w_i \cdot x_i}{b},\ a \geq 0,\ b > 0,\ w_i > 0\ \forall i. \tag{3.2}$$

WLC contains the special cases of the weighted sum ($a = 0,\ b = 1$), the weighted average ($a = 0,\ b = \sum w_i$), and the arithmetic mean ($a = 0,\ b = n,\ w_i = 1\ \forall i$). WLC strategies are injective, idempotent, and allow for prioritisation through weighting. However, these strategies do not allow for dependencies and do not satisfy the weakest link principle.

A related set of strategies are the weighted product (WP) strategies:

$$S_{WP}(\mathbf{x}) = a + b \cdot \prod_{i=1}^{n} x_i^{w_i},\ a \geq 0,\ b > 0,\ w_i \in (0, 1]\ \forall i. \tag{3.3}$$

Among the WP strategies are the simple product ($a = 0,\ b = 1,\ w_i = 1\ \forall i$) and the geometric mean ($a = 0,\ b = 1,\ w_i = \frac{1}{n}\ \forall i$). WP strategies satisfy the same properties as WLC strategies, except for the idempotence property which these strategies do not satisfy.

Using the weighted maximum (WM) - $S_{WM}(\mathbf{x}) = max\{w_1 \cdot x_1, \ldots, w_n \cdot x_n\},\ w_i > 0\ \forall i$ - metric value as the aggregated value is uncommon in most disciplines, since this strategy is not injective. However, it is used in the security field (Lippmann, Riordan, et al., 2012), and is in fact an extreme case of satisfying the weakest link principle. WM allows for prioritisation, although the basic maximum function does not.

The complementary product is another aggregation strategy that is uncommon outside of the security field (Lippmann, Riordan, et al., 2012). Let $\hat{x}_i$, for $i = 1, 2, \ldots, n$, denote the metric value normalised to $[0, 1)$. Let w_i be the weight of metric i for $i = 1, 2, \ldots, n$. We define the weighted complementary product (WCP) class as:

$$S_{WCP}(\mathbf{x}) = a \cdot \left(1 - \prod_{i=1}^{n}(1 - \hat{x}_i)^{w_i}\right),\ a > 0,\ w_i \in (0, 1]\ \forall i. \tag{3.4}$$

The regular complementary product is achieved with $a = 1$ and $w_i = 1\ \forall i$. WCP strategies are injective and can satisfy the prioritisation and weakest link principles, depending on the values of w_i.

None of the strategies considered so far consider dependency. Bayesian networks (BN) are probabilistic graphical models, often of a causal nature, that are commonly applied in the security field (Kordy et al., 2014). In BN aggregation strategies, the metric values x_i are assumed to originate from discrete, bounded random variables X_i, corresponding to the metrics $i = 1, \ldots, n$. The conditional dependencies between the random variables, and with a potential unobserved variable Y, are made explicit. This allows us to infer the probabilities of different values of Y, based on the metric values x_i. BN strategies are injective, but not idempotent. Although prioritisation is generally not a goal within these strategies, the prioritisation property will usually be satisfied. BN strategies accommodate dependencies by their nature, but will mostly not satisfy the weakest link principle.

The strategy classes presented in Table 3.2 are not exhaustive but do cover the large majority of all aggregation strategies employed, as we show in Section 3.4. Two examples of other possibilities are the use of analytic network process (ANP) techniques (Brožová et al., 2016; Lo and W.-J. Chen, 2012), which relate to the deterministic equivalent of Bayesian networks, and the analysis of game-theoretic equilibria (Rass et al., 2017). What is common to all strategies, is that none satisfy all criteria of Table 3.2, where we should additionally note that strategies within the classes of weighted maximum and weighted complementary product cannot satisfy the prioritisation and weakest link properties at the same time.

Table 3.2: Various classes of metric aggregation strategies, and important security-related properties their strategies can possess.

AGGREGATION	INJECTIVE	IDEMPOTENT	PRIORITISATION	DEPENDENCE	WEAKEST LINK
Weighted linear combination	✓	✓	✓	×	×
Weighted product	✓	×	✓	×	×
Weighted maximum	×	✓	✓	×	✓
Weighted complementary product	✓	×	✓	×	✓
Bayesian network	✓	×	✓	✓	×

3.2.4 *Adaptability*

Adaptability is crucial to any cybersecurity solution (Evesti and Ovaska, 2013). Especially when measuring cybersecurity, a rigid solution that does not adapt to a changing environment or a new use case is far from optimal (Baars et al., 2016). It is not surprising to see, then, that adaptability is a key focus of many studies (de las Cuevas et al., 2015; Yigit Ozkan, Spruit, et al., 2019), although operationalisation of adaptability is still a challenge (Evesti and Ovaska, 2013).

We consider adaptability to be "the state of being able to change to work or fit better" (J.-H. Cho et al., 2019). This definition outlines two important dimensions of adaptability. Firstly, a solution is considered adaptable if it can change to work better. There are several reasons why a cybersecurity metric solution may not be functioning as it should. This can relate to problems with the metrics themselves, such as missing or dirty data (W. Kim et al., 2003). It can also relate to a changing security landscape, that invalidates an existing model. This phenomenon is known as concept drift (Widmer and Kubat, 1996). Secondly, a solution is considered adaptable if it can change to fit better. Generally, cybersecurity solutions in research are made to fit their use case. We can determine their adaptability in the 'fitting' dimension by determining how easily the solution can be deployed at other (similar) use cases.

Adaptability is significant in the SME context. The SME landscape is diverse (European DIGITAL SME Alliance, 2020), and SMEs often lack the knowledge and expertise to perform extensive adaptations independently (Shojaifar, Fricker, and Gwerder, 2020). In Section 3.6, we assimilate observations from earlier research and our results of Section 3.4 to provide suggestions for improving solution adaptability.

3.3 SYSTEMATIC REVIEW METHODOLOGY

We performed a systematic literature review to address our research questions. To ensure broad coverage of the cybersecurity metrics field, we employed a novel Systematic Review Methodology Blending Active Learning and Snowballing (SYMBALS, (van Haastrecht, Sarhan, Yigit Ozkan, et al., 2021)), which

combines existing methods into a swift and accessible methodology, while following authoritative systematic review guidelines (Kitchenham and Charters, 2007; Liberati et al., 2009; Moher et al., 2015).

Active learning is one of the cornerstones of the SYMBALS approach. Active learning is commonly applied in the title and abstract screening phase of systematic reviews, where researchers start with a large set of papers and prefer to not screen them all manually (van de Schoot et al., 2021). Active learning is uniquely suited to this task, as this machine learning method selects the ideal data points for an algorithm to learn from.

SYMBALS complements active learning with backward snowballing. From a set of included papers, a researcher can find additional relevant papers by consulting references (backward snowballing) and citations (forward snowballing) (Wohlin, 2014). Snowballing has proven to be a valuable addition to systematic reviews, even when reviews already include an extensive database search (Mourão, Pimentel, et al., 2020). Backward snowballing is especially useful in uncovering older relevant research. Forward snowballing is not employed within SYMBALS, based on the observation that databases generally have excellent coverage of recent peer-reviewed research.

After the development and evaluation of a systematic review protocol for this research, we commenced with the database search step of SYMBALS. We retrieved research from abstract databases (Scopus, Web of Science) and full-text databases (ACM Digital Library, IEEE Xplore, PubMed Central).

The Scopus API was used to retrieve an initial set of relevant research. Results from other sources were then successively added to this set. The order in which sources were consulted can be surmised from Table 3.3. The Python Scopus API wrapper 'pybliometrics' (Rose and Kitchin, 2019) was used to retrieve all research available through the Scopus API, that satisfied the query:

```
AUTHKEY((security* OR cyber*)
AND (assess* OR evaluat* OR measur* OR metric* OR model* OR risk*
    OR scor*))
AND LANGUAGE(english) AND DOCTYPE(ar OR bk OR ch OR cp OR cr OR re
    )
```

The 'AUTHKEY' field corresponds to the keywords that authors provided for a paper. Our search query is intentionally broad, as the SYMBALS methodology allows us to deal with larger quantities of research, and we aim to exclude as little relevant research as possible at this stage. We did choose to only include English language research and document types where extensive and verifiable motivations for findings can be reported.

Table 3.3 summarises the query results. ACM Digital Library and IEEE Xplore limit the number of accessible papers to 2,000. This means only the 2,000 most relevant papers from these sources could be considered. Moreover, IEEE Xplore only allows the use of 6 wildcards in the search query. We removed the 'security' and 'cyber' wildcards for the IEEE Xplore search to comply with this limitation. Any research without an abstract was excluded,

Table 3.3: Statistics regarding the different databases used in the search procedure.

SOURCE	RESULTS	UNIQUE
Scopus	21,964	21,964
Web of Science	7,889	1,782
ACM Digital Library	2,000	660
IEEE Xplore	2,000	1,256
PubMed Central	660	111
Total	**34,513**	**25,773**

as this is vital to the active learning phase of SYMBALS. This led to a small set of exclusions from the PubMed Central database. Duplicate removal was performed based on the research title, although we found that this process was not perfect, due to different character sets being accepted in different databases.

Altogether, our dataset resulting from database search comprised 25,773 papers. This exemplifies the broad scope of our research, as the largest initial set of papers from the reviews in Table 3.1 comprised 4,818 papers (Morrison et al., 2018).

The set of 25,773 papers is too large to perform data extraction directly. This is where the active learning phase of SYMBALS comes in. We chose to use ASReview in this phase, a tool that offers active learning capabilities for systematic reviews, specifically for the title and abstract screening step (van de Schoot et al., 2021). Many other active learning tools exist that are worth considering (Harrison et al., 2020). However, we found ASReview effective and easy to use, and additionally value the commitment its developers have made to open science. This shows in, among other things, the codebase that they made available open-source.

In the ASReview process, as well as in the later review phases, we made use of the following inclusion and exclusion criteria:

- Inclusion criteria:
  - I1: The research concerns cybersecurity metrics and discusses how these metrics can be used to assess the security of a (hypothetical) cyber-system.
  - I2: The research is a review of relevant papers.
- Exclusion criteria:
  - E1: The research does not concern cyber-systems.
  - E2: The research does not describe a concrete path towards calculating cybersecurity metrics (only applied if I2 is not applicable).
  - E3: The research has been retracted.
  - E4: There is a more relevant version of the research that is included.

- E5: The research was automatically excluded due to its assessed irrelevance by the ASReview tool.
- E6: The research does not satisfy the database query criteria on language and document type.
- E7: No full-text version of the research can be obtained.
- E8: The research is of insufficient quality.
- E9: The research does not contain at least one socio-technical cybersecurity metric.

Exclusion criterion E8 relates to the quality assessment phase of SYMBALS, which is explained below. Criterion E9 requires the consideration of the full text to be determined, as abstracts do not contain enough information to make a decision regarding this intricate topic (Brereton et al., 2007). Thus, neither of these criteria were applied during title and abstract screening.

ASReview requires users to specify prior relevant and irrelevant papers to train its algorithm. We used five papers as initial indications of relevance to ASReview (Allodi and Massacci, 2017; J.-H. Cho et al., 2019; Noel and Jajodia, 2014; Spruit and Röling, 2014; Stolfo et al., 2011). These papers were chosen since they cover diverse topics, were written by different authors at different times and were published in different journals and conferences. ASReview additionally provides the option to label a certain number of random papers before proceeding, assuming that a significant proportion of these papers will be irrelevant. This provides the algorithm with a balance of relevant and irrelevant papers for training. We labelled 5 random papers, giving us a total training set of 10 papers.

The ASReview tool then presents the paper whose classification it deems most informative to learn from. The tool quickly learns to distinguish between relevant and irrelevant papers. By presenting the researcher mostly relevant papers, the process of discovering relevant papers is accelerated.

Although ASReview offers several classifier options, we employed the default Naïve Bayes classifier using term frequency-inverse document frequency (TF-IDF) feature extraction and certainty-based sampling. The default settings have been shown to produce consistently good results and are additionally commonly available in other active learning tools (van de Schoot et al., 2021). Thus, our decision to use the default settings can be motivated both from a performance and a reproducibility standpoint.

At some point in the active learning process, mostly irrelevant research remains. To reduce the time spent on assessing irrelevant research, a stopping criterion is used (van de Schoot et al., 2021). We stop evaluating research when the last 20 reviewed papers were considered irrelevant, although more sophisticated stopping criteria exist that are worth considering (Cormack and Grossman, 2016). All research that was not evaluated at this stage, was

Table 3.4: The quality criteria applied to 60 papers during the quality assessment phase. Possible responses were strongly disagree (SD), disagree (D), neutral (N), agree (A), or strongly agree (SA).

ASPECT	CRITERION	SD	D	N	A	SA
Reporting	There is a clear statement of the research aims.	0	4	7	28	21
	There is an adequate description of the research context.	0	6	11	17	26
	The paper is based on research.	0	3	3	16	38
Rigour	Metrics used in the study are clearly defined.	0	10	19	16	15
	Metrics are adequately measured and validated.	1	24	22	8	5
	The data analysis is sufficiently rigorous.	0	21	17	14	8
Credibility	Findings are clearly stated and related to research aims.	0	8	19	25	8
	Limitations and threats to validity are adequately discussed.	30	18	8	2	2
Relevance	The study is of value to research and/or practice.	0	9	12	28	11

excluded based on exclusion criterion E5. As Figure 3.1 shows, 1,644 papers remained after the active learning phase.

Figure 3.1: Visualisation of the SYMBALS steps as applied in our cybersecurity metric systematic review.

We then proceeded with the backward snowballing phase of SYMBALS. We followed the ASReview evaluation order in our backward snowballing procedure. We concluded backward snowballing once 10 consecutive papers contained no new references satisfying the inclusion criteria. As can be seen in Figure 3.1, 1,796 papers were contained in our inclusion set after the completion of this phase.

SYMBALS specifies quality assessment as an optional step, but given the large number of papers remaining, assessing quality was deemed necessary. Table 3.4 outlines the quality criteria that were applied. Commonly used research quality criteria were adapted for use with a Likert scale (Y. Zhou et al., 2015). Statements could be responded to with strongly disagree, disagree, neutral, agree, or strongly agree. Instead of applying these criteria to all 1,796 inclusions, the two researchers involved in quality assessment evaluated 40 papers, with 20 papers being evaluated by both researchers.

A simple, yet effective, solution to extrapolate these results is to train a binary decision tree on basic research characteristics, to create a model that can distinguish research of sufficient quality from research of insufficient quality. The five Likert scale responses were assigned scores of 0 (strongly disagree), 0.25 (disagree), 0.5 (neutral), 0.75 (agree), and 1 (strongly agree). Summing the

quality criteria scores, each paper received a score between 0 and 9. To make the problem a binary decision problem, we labelled papers with a score of at least 6 as having sufficient quality. The height of this threshold determines how strict the eventual model will be.

Next, we split our set of 60 evaluated papers into a training set of 48 papers (80%) and a test set of 12 papers (20%). To be able to train a model on this set, we need explanatory variables which explain the quality scores obtained by the papers. We opted to use three features: years since publication, citation count, and the number of pages. The maximum depth of the binary decision tree was set to 3, meaning at most 3 binary splits are performed before classifying a paper as having sufficient or insufficient quality. The model was trained on the 48 training papers and evaluated on the 12 test papers. Despite - or perhaps because of - the model's simplicity, 11 of the 12 test papers were labelled correctly. The only incorrect labelling occurred in an edge case with a quality score of 6. Similar results were obtained in replications with different random seeds. Figure 3.1 shows that 516 papers remained after applying the binary decision tree to our complete inclusion set.

Finally, we applied exclusion criterion E9 using a manual screening process, to filter out the papers that do not consider the social side of cybersecurity, as defined in Section 3.5. Figure 3.1 shows that in total 60 papers were included after our filtering step.

3.4 RESULTS

In this section, we focus on descriptive analysis of aggregate results. In Sections 3.5 and 3.6, we will dive deeper, to interpret and contextualise the results.

Figure 3.2 depicts the relative prevalence of each of the five ADKAR factors over the years. Since 2010, awareness and reinforcement together constituted over half of the ADKAR considerations. Desire is the element that receives the least attention in research. Table 3.5 lists the related concepts which we encountered and mapped to each of the ADKAR terms.

Part of the reason for the prevalence of reinforcement research is that cybersecurity training and education belong to this ADKAR element. Researchers feel that organisational reinforcement is an important aspect of the social side of cybersecurity. At the same time, reinforcement can be easier to measure than other factors, which may offer a partial explanation for its prevalence. For example, many researchers choose to include a metric of cybersecurity awareness training (reinforcement), rather than of cybersecurity awareness itself (awareness).

Various security concepts were assessed in our inclusions, as shown in Table 3.6. Some researchers choose to measure security itself (Bhilare et al., 2008; You et al., 2015), but this approach is too general for most. Risk was assessed in two-thirds of all papers. This is interesting, as risk can be seen as having

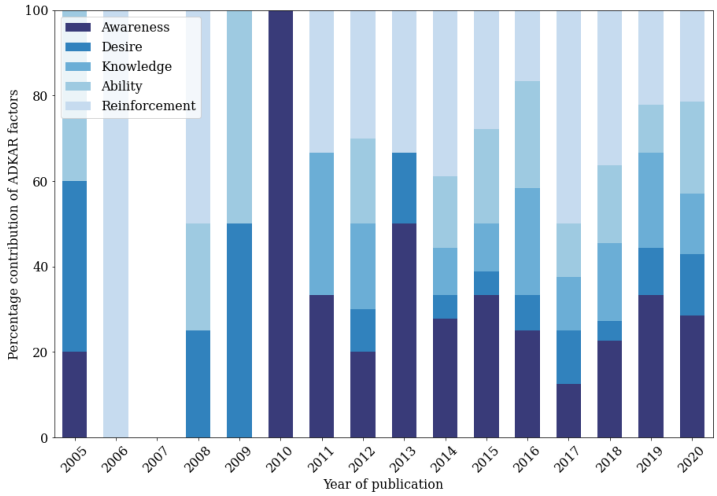


Figure 3.2: The consideration of the five ADKAR factors over the years, based on the 60 inclusions of our systematic review.

Table 3.5: The ADKAR factors and the related concepts we encountered which were associated to each factor.

ADKAR	ABBREVIATION	RELATED CONCEPTS
Awareness	AW	Consciousness
Desire	DE	Motivation, loyalty, attendance
Knowledge	KN	Understanding
Ability	AB	Behaviour, capability, capacity, experience, skill
Reinforcement	RE	Culture, education, evaluation, policy, training

Table 3.6: The various security assessment concepts discussed in research, with an indication of the ADKAR elements covered and the aggregation strategies employed. Each paper should consider at least one ADKAR element. A paper may not aggregate at all, but could also employ several aggregation strategies. Reviews were not labelled with a specific assessment concept.

		ADKAR elements					Aggregation strategy classes					
CONCEPT	TOTAL	AW	DE	KN	AB	RE	WLC	WP	WM	WCP	BN	NONE
Risk	40	24	9	14	19	28	27	10	7	1	4	4
Awareness	5	5	3	4	3	2	3	1	1	0	0	2
Maturity	5	0	0	0	0	5	4	0	1	0	0	0
Resilience	3	3	1	0	1	1	3	0	0	0	0	0
Security	2	1	0	0	0	2	1	0	0	0	0	1
Vulnerability	1	1	0	1	0	0	1	0	0	0	0	0

a negative connotation, whereas awareness, maturity, and resilience have positive connotations. This finding conflicts with the general tendency in the security community to favour SDT approaches over the fear- and threat-based approaches more associated with PMT (Menard et al., 2017), especially in the context of organisations (N. Yang et al., 2020).

When analysing the ADKAR factors by assessment concept, the papers assessing security maturity stand out. These papers place a large focus on the organisational reinforcement of security and ignore all other ADKAR factors. This is not a surprising finding. Maturity is generally a concept that requires an assessment of the organisation, rather than the individuals who make up this organisation.

Table 3.6 shows that most papers stick to WLC, WP, and WM as aggregation strategies. It is worth pointing out that not aggregating is a reasonable choice. If it is not necessary for a particular context, it should be avoided, based on our conclusion from Table 3.2 that no aggregation method satisfies all ideal security properties.

Table 3.7 focuses on the actors that were considered from the social viewpoint. Almost all papers focus solely on the defender. It is interesting to see that the desire and ability factors of ADKAR are much more prominent in research including the attacker. We would expect to see more focus from research on desire, and the related concept of motivation, based on the important role that motivation and internalisation play in SDT and PMT (Padayachee, 2012). Desire and motivation are not easily measurable concepts, but metrics such as 'attendance at security sessions' can serve as useful proxies here (Manifavas et al., 2014).

Nearly all research that considers the attacker perspective considers the real-life threat environment as specified in Gollmann et al. (2015). In papers covering the defender, it is quite common to ignore threats entirely (Y. Shin et al., 2011) or to use a proxy such as the prevalence of vulnerabilities to

Table 3.7: The different social viewpoints considered in our inclusions.

		ADKAR					
SOCIAL VIEWPOINT	TOTAL	AW	DE	KN	AB	RE	REAL-LIFE THREAT
Defender	52	33	7	17	17	37	18
Attacker	5	0	4	1	5	0	5
Both	3	2	3	1	3	3	2

Table 3.8: Different aggregation strategy classes and the situations in which they were employed.

	Classification		
AGGREGATION STRATEGY	THEORETICAL	IMPLEMENTATION	REVIEW
WLC	38	1	3
WP	11	0	0
WM	8	1	0
WCP	1	0	0
BN	4	0	0
None	7	2	1

represent threats (Marconato et al., 2013). This is remarkable given the vital role that threat perception plays in both SDT and PMT (Menard et al., 2017).

Table 3.8 groups research based on the employed aggregation strategy. Inclusions were classified into one of three classes: theoretical, implementation, or review. The research was classified as an implementation if either clear and described actions were taken based on the implemented method, or the model was assessed at more than one point in time. This strict requirement explains why most papers were classed as theoretical.

One immediately notices from Table 3.8 that two of the four implementation papers do not employ an aggregation strategy. As we discussed in Section 3.2.3 and showed in Table 3.2, aggregation should only be carried out if deemed necessary. In half of the implementation research of our inclusions, researchers felt the benefits of aggregation did not outweigh the drawbacks.

We additionally see that most research sticks to WLC and WP strategies, which do not satisfy the weakest link principle and cannot take into account dependencies. Researchers prefer simple and explainable strategies, that are injective or idempotent, over strategies that satisfy more security properties. Out of our 60 inclusions, 10 used fuzzy logic approaches. Although translating qualitative statements to fuzzy numbers differentiates these methods from approaches using crisp numbers, most still use some combination of WLC, WP, and WM to aggregate (for example, (X. Li et al., 2018; Shameli-Sendi, Shajari, et al., 2012; Silva et al., 2014)).

Exceptions are Lo and W.-J. Chen (2012) and Brožová et al. (2016), who use an ANP approach to capture dependencies. Lo and W.-J. Chen (2012), Brožová et al. (2016) and the four papers using a bayesian network approach (Dantu

Table 3.9: ADKAR and aggregation strategy frequencies of enterprise research and other research.

		Application area		
PROPERTY	VALUES	ANY ENTERPRISE	M/L ENTERPRISE	OTHER
ADKAR	AW	9	6	20
	DE	3	1	10
	KN	7	2	10
	AB	6	3	16
	RE	11	13	15
Aggregation	WLC	13	7	22
	WP	0	3	8
	WM	2	2	5
	WCP	0	0	1
	BN	0	1	3
	None	1	4	5

and Kolan, 2005; Dantu, Kolan, and Cangussu, 2009; N. Feng et al., 2014; Sahinoglu, 2008) are the only papers that consider dependencies between metrics. Interestingly, all of these papers were published in 2016 or earlier. It is not immediately clear what the underlying reason is for the current drought in research considering dependencies, but it is certainly a research area that deserves more attention.

Table 3.9 provides detailed results regarding the research application area. Although more enterprise sizes were considered, we only encountered research applicable to medium- and large-sized enterprises, and research applicable to any enterprise size. As with research focused on maturity modelling, we see a strong focus on the reinforcement factor of ADKAR in enterprise research, especially for larger enterprises.

In research intended to apply to any enterprise, Table 3.9 shows that WLC is by far the most popular aggregation strategy class. The only other strategy class that is used is WM. We believe it is not a coincidence that these are the only aggregation strategy classes that are both injective and idempotent. Strategies with these properties are likely to be more intuitive and easy to understand, as explained in Section 3.2.3. Therefore, it is not surprising that these strategies are proposed in research addressing all enterprise sizes, since especially smaller businesses need to be motivated through approachable solutions.

Regarding adaptability, of the 56 inclusions that were not review papers, 44 do not make any consideration for missing or dirty data. Of the papers that do consider one or both of these issues, the most common strategy is to ignore the associated problems. Out of these 56 papers, 46 are not able to adapt to a security event occurring, mostly since they do not operate in a live setting, but are formulated as periodic assessments. Even then, most authors do not cover this topic, and it is certainly not always clear how the security assessment would be adapted after an incident.

Concept drift and adaptation to other use cases are also often not considered. Just four of our inclusions explicitly consider concept drift and no paper mentions a concrete timeline for when a solution should be updated. Adaptation to other use cases is discussed in 24 of our inclusions. However, the majority of these papers only give a rough outline of how the solution could be adapted. A better practice would be to give concrete guidelines on how to adapt the solution or to immediately analyse several use cases. The former approach was not seen in research, whereas the latter was (for example, (Chan, 2011; M.-K. Chen and S.-C. Wang, 2010; Luh et al., 2020; Proença and Borbinha, 2018)).

3.5 SOCIO-TECHNICAL CYBERSECURITY FRAMEWORK FOR SMES

To offer more insight into how we can create effective cybersecurity assessment solutions for SMEs, we position our results and findings in the STS analysis framework of Davis et al. (2014). Figure 3.3 shows the view of STS as consisting of six internal social and technical aspects, within an external environment. We rename the 'Buildings/Infrastructure' aspect of Davis et al. (2014) to 'Assets.' This ensures that our view is better aligned with standard terminology in cybersecurity literature. Based on the importance of policies in socio-technical cybersecurity frameworks (Malatji, Von Solms, et al., 2019), we explicitly include policies in the 'Processes/Procedures' aspect of Davis et al. (2014) and rename this aspect to 'Processes.'

The socio-technical system we study is the SME, in the context of cybersecurity. However, the complete set of SMEs is too diverse to consider this group as a single collective. This is why the European DIGITAL SME Alliance proposes to use four SME categories, based on the different roles SMEs can play in the digital ecosystem: start-ups, digitally dependent SMEs, digitally based SMEs, and digital enablers (European DIGITAL SME Alliance, 2020). The European DIGITAL SME Alliance specifies these categories in the context of cybersecurity standardisation, which is intricately related to our cybersecurity assessment setting, making it a suitable classification.

The European DIGITAL SME Alliance defines start-ups as SMEs where "security has a low priority." They "typically neglect (or are not aware of) requirements" for running a secure business. Digitally dependent SMEs are companies that depend on digital solutions (as end users) to run their business. Digitally based SMEs "highly depend on digital solutions for their business model," and, finally, digital enablers are SMEs that develop and provide digital solutions (European DIGITAL SME Alliance, 2020).

Table 3.10 introduces our framework, which synthesises the SME categories of the European DIGITAL SME Alliance (2020) with the STS aspects of Davis et al. (2014). Each SME category has different cybersecurity goals based on their different roles in the digital ecosystem. In Table 3.10, the SME categories are ordered from least to most mature regarding cybersecurity. We expect the

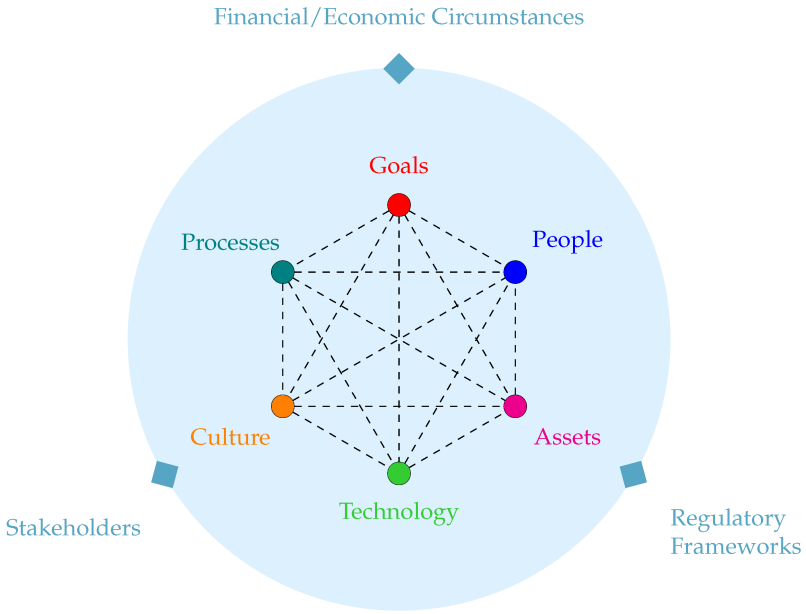


Figure 3.3: A socio-technical system embedded within an external environment, based on Davis et al. (2014).

more mature SME categories to have achieved the goals of less mature SME categories.

Our framework was constructed based on earlier cybersecurity frameworks focusing on SMEs (Benz and Chatterjee, 2020; Carías, Borges, et al., 2020; Cholez and Girard, 2014) or STS (AlHogail, 2015; Da Veiga et al., 2020; Malatji, Marnewick, et al., 2020; Malatji, Von Solms, et al., 2019; Sittig and H. Singh, 2016). Interestingly, none of these frameworks focused on both SMEs and STS. To address the singular characteristics of our setting, we additionally incorporated the findings from our systematic review, as well as principles for designing cybersecurity maturity models for SMEs (Yigit Ozkan and Spruit, 2020), in our framework. Our findings appear most prominently in the 'Technology' aspect, explaining why this column of Table 3.10 contains relatively few references to earlier work.

Our results relating to the various ADKAR dimensions serve as input for the 'People' and 'Culture' aspects. Start-ups and digitally dependent SMEs should focus on making their employees aware and providing initial cybersecurity knowledge to inspire desire and motivation. This can be achieved through a culture of organisational commitment to cybersecurity (AlHogail, 2015; Da Veiga et al., 2020). Digitally based SMEs and digital enablers should progress through the ADKAR phases, with the aid of cybersecurity training, policy, and assessment. Eventually, employees should mutually reinforce each

Table 3.10: Socio-technical cybersecurity framework for SMEs.

	Socio-Technical Aspects					
SME CATEGORY	GOALS	PEOPLE	CULTURE	PROCESSES	TECHNOLOGY	ASSETS
Start-ups	Realise cybersecurity necessity (Malatji, Von Solms, et al., 2019) due to external environment factors. Move from a non-existent cybersecurity culture to initial, informal cybersecurity measures (Benz and Chatterjee, 2020; Malatji, Marnewick, et al., 2020; Malatji, Von Solms, et al., 2019).	Define training plans and start creating cybersecurity awareness (Carías, Borges, et al., 2020).	Initial cybersecurity policies and procedures show management commitment, ensuring employee support (AlHogail, 2015; Da Veiga et al., 2020).	No standardised processes yet (Malatji, Von Solms, et al., 2019). SME gains awareness on cybersecurity policies, processes, procedures, standards and regulation.	Employ a threat-based risk assessment tool requiring no knowledge of SME assets, using no/intuitive aggregation. External support to understand and implement countermeasures.	Understand relevant and critical cybersecurity asset types (Carías, Borges, et al., 2020).
Digitally dependent	Start formalising cybersecurity processes. Define, manage, and communicate cybersecurity strategy (Benz and Chatterjee, 2020; Carías, Borges, et al., 2020; Malatji, Marnewick, et al., 2020; Malatji, Von Solms, et al., 2019).	Continue building awareness (Da Veiga et al., 2020). Stimulate desire through knowledge acquisition (AlHogail, 2015). Evaluate gaps in ability (Carías, Borges, et al., 2020).	Management support and cybersecurity trainings stimulate employees (Da Veiga et al., 2020) and change their perception (AlHogail, 2015).	Formulate basic (reactive) cybersecurity policies, processes, and procedures (Da Veiga et al., 2020; Malatji, Von Solms, et al., 2019); likely not yet universally applied across business units (Malatji, Von Solms, et al., 2019).	Employ a threat-based risk assessment tool using no/intuitive aggregation. External support to implement countermeasures.	Systematically identify and document relevant assets and their baseline configurations (Carías, Borges, et al., 2020).
Digitally based	Establish a formal cybersecurity programme that facilitates continuous improvement and compliance with regulation (Benz and Chatterjee, 2020; Carías, Borges, et al., 2020; Malatji, Marnewick, et al., 2020; Malatji, Von Solms, et al., 2019).	Advance cybersecurity knowledge and ability through clearly communicated and documented trainings (Carías, Borges, et al., 2020; Da Veiga et al., 2020; Malatji, Von Solms, et al., 2019).	Regular communication and education (Da Veiga et al., 2020), backed by rewards and deterrents (AlHogail, 2015), ensures secure employee behaviour (AlHogail, 2015; Da Veiga et al., 2020).	Processes defined and documented proactively, communicated via awareness and training sessions (Da Veiga et al., 2020; Malatji, Von Solms, et al., 2019). Information sharing agreements defined (Carías, Borges, et al., 2020).	Use a risk assessment framework or maturity model with adequately motivated aggregation. Implement basic countermeasures (Carías, Borges, et al., 2020), external support for complex countermeasures.	Manage asset changes and periodically maintain assets (Carías, Borges, et al., 2020).
Digital enablers	Embed and automate cybersecurity processes (Benz and Chatterjee, 2020; Cholez and Girard, 2014; Malatji, Marnewick, et al., 2020; Malatji, Von Solms, et al., 2019), which, combined with collaborative stakeholder relationships (Carías, Borges, et al., 2020), promote internal and external trust in the SME cybersecurity posture (Da Veiga et al., 2020).	Employees mutually reinforce their cybersecurity abilities, possibly captured in official cybersecurity roles (Da Veiga et al., 2020).	Regular evaluations (Cholez and Girard, 2014; Malatji, Von Solms, et al., 2019) stimulate naturally secure behaviour (Da Veiga et al., 2020), where national culture and regulations are recognised (AlHogail, 2015). An environment of trust with stakeholders exists (Carías, Borges, et al., 2020; Da Veiga et al., 2020).	Successive comparisons of assessment results facilitate continuous process improvement (Cholez and Girard, 2014; Malatji, Von Solms, et al., 2019). Business continuity plan defined and communicated to external stakeholders (Carías, Borges, et al., 2020).	Use a risk assessment framework or maturity model with advanced aggregation. Independently implement countermeasures (Carías, Borges, et al., 2020) and actively detect anomalies (Carías, Borges, et al., 2020), with the help of automated tools (Malatji, Von Solms, et al., 2019).	Identify and document internal and external dependencies of assets, to help in determining the SME attack surface. Actively monitor assets (Carías, Borges, et al., 2020).

other's cybersecurity abilities (Da Veiga et al., 2020). The ideal cybersecurity culture will lead to trust from both the people inside the SME, as well as the environment outside of the SME (Carías, Borges, et al., 2020; Da Veiga et al., 2020).

Start-ups and digitally dependent SMEs are often not aware of the existence of cybersecurity standards (European DIGITAL SME Alliance, 2020). These SMEs should first become aware and then begin to formulate basic cybersecurity policies, processes, and procedures (Da Veiga et al., 2020; Malatji, Von Solms, et al., 2019). Digitally based SMEs should have formal processes in place to reinforce desired cybersecurity behaviour of employees (Malatji, Von Solms, et al., 2019). Digital enabler SMEs should strive towards continuous process improvement (Cholez and Girard, 2014; Malatji, Von Solms, et al., 2019), which enables business continuity (Carías, Borges, et al., 2020).

We map the 'Technology' aspect of STS to the advised cybersecurity assessment approach and tooling for the SME. This is in line with the approach of Malatji, Von Solms, et al. (2019), who incorporate "cybersecurity tools and resources" in the 'Technology' aspect of their socio-technical cybersecurity framework.

Start-ups should understand relevant cybersecurity asset types and digitally dependent SMEs should begin identifying and documenting assets (Carías, Borges, et al., 2020). Without an asset inventory or internal cybersecurity expertise, most risk assessment and maturity model approaches are not suited to these SMEs. Additionally, they are just beginning to cultivate a desire among employees to improve cybersecurity. Incorporating the real-life threat environment (Gollmann et al., 2015) is an attractive option to promote motivation. Focusing on the real-life threat environment can increase the feelings of task relevance and significance employees feel, which are key motivators (Kam et al., 2020). This is why we advise a threat-based cybersecurity risk assessment approach for start-ups and digitally dependent SMEs.

In the same vein, we advise to not aggregate scores in cybersecurity assessment solutions for start-ups and digitally dependent SMEs. If aggregation is deemed necessary, injective and idempotent aggregation strategies should be used, such as WLC and WM. Strategies that satisfy injectivity and idempotence can be seen as intuitive. Using these strategies allows for feelings of competence and relatedness among employees, which stimulate motivation (Menard et al., 2017). This puts employees in a position to be a part of the solution to SME cybersecurity challenges, rather than being the source of the challenges (Zimmermann and Renaud, 2019).

The combination of simple aggregation and a threat-based approach offers another benefit: the corresponding assessments do not necessarily require extensive internal expertise and data. Many of the more complex aggregation strategies and comprehensive assessment approaches require cybersecurity experts at the SME to determine parameters and weights. Such resources are

limited at SMEs (Heidt et al., 2019), and especially at start-ups and digitally dependent SMEs. This is why assessment approaches for these SMEs should preferably be largely based on data that can be automatically collected. Threat-based approaches are ideally suited to this requirement, as general incident data is widely available (Y. Liu et al., 2015), and can be mapped to threats to offer SMEs insight into what is important for them (Casola et al., 2019).

Digitally based SMEs and digital enablers can be expected to have a complete inventory of assets (Carías, Borges, et al., 2020). Digital enablers should additionally be aware of internal and external dependencies (Carías, Borges, et al., 2020), allowing them to specify their attack surface (Manadhata and Wing, 2011). For these SME categories, complete risk- and maturity assessments are desirable. Digital enablers will often require comprehensive assessments that can prove compliance with cybersecurity standards and regulations.

Digitally based SMEs should consider using aggregation strategies that reflect desirable security properties, such as the weakest link principle. Using a WCP strategy can guide these SMEs towards more accurate assessments, although intuitiveness is sacrificed. Digital enablers with cybersecurity expertise, a specified attack surface, and large volumes of internal data, should consider more advanced aggregation strategies.

Figure 3.4 provides a visual summary of the STS interactions inherent to our framework. We use coloured arrows to indicate interactions that are explicitly mentioned in Table 3.10. It is implicit in the STS model of Davis et al. (2014) that all aspects are interrelated.

The direction of the arrows indicates which aspect serves as an input for another aspect. For start-ups, the external environment aspects motivate the SME to realise the necessity of investing in cybersecurity, leading to the initial goals. For digitally dependent SMEs, the goals formulated by management serve as catalysts for culture and processes. We observe that from an initial external motivation for start-ups, SMEs gradually build up internal interactions. For digital enablers, we see many interactions, both internally and with the external environment.

3.6 DISCUSSION

We extensively analysed and interpreted our results in Sections 3.4 and 3.5. This section will focus on a discussion of our research questions and the potential limitations of our research.

Our first research question asked: how are cybersecurity metrics aggregated in socio-technical cybersecurity measurement solutions? One interesting finding from Table 3.8 is that half of the research involving implementations did not aggregate at all. Table 3.2 gives a partial explanation for this phenomenon: no aggregation strategy satisfies all desirable security properties. Thus, aggregation should preferably be avoided. Nevertheless, aggregation using basic approaches such as WLC is prevalent, with 42 of our 60 inclusions

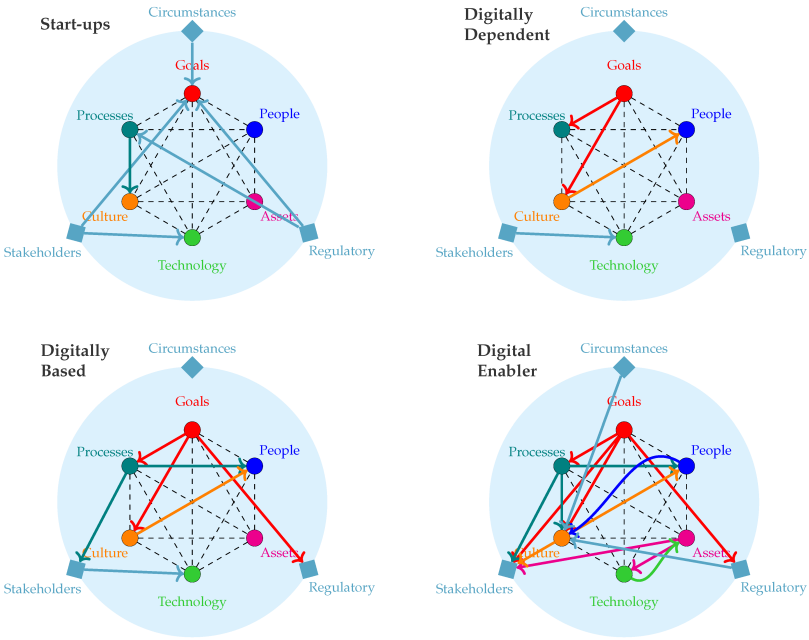


Figure 3.4: A visualisation of the framework presented in Table 3.10 using the representation of Figure 3.3.

using this aggregation technique. We observed a clear lack of dependency consideration among metrics, which could be solved using Bayesian network (Dantu and Kolan, 2005; Dantu, Kolan, and Cangussu, 2009; N. Feng et al., 2014; Sahinoglu, 2008) or ANP techniques (Brožová et al., 2016; Lo and W.-J. Chen, 2012). Our cybersecurity framework presented in Table 3.10 provides clear guidance on which aggregation strategies suit which SME categories.

Our second research question was formulated as: How do aggregation strategies differ in cybersecurity measurement solutions relevant to SMEs and all other solutions? Our analysis of Table 3.9 demonstrated that in enterprise research little to no attention is paid to aggregation strategies that satisfy the weakest link and dependency properties. One of the main obstacles in making aggregation strategies suitable for SMEs is the time and expertise required to carry them out. Generally, more complex aggregation strategies require the determination of more parameters and relationships, which in turn often requires consultation of security experts at the cyber-system being assessed (for example, (Alencar Rigon et al., 2014; Damenu and Beaumont, 2017; Proença and Borbinha, 2018; Shokouhyar et al., 2018)). This expertise is rarely available at smaller SMEs, although when it is, ANP approaches (Brožová et al., 2016; Lo and W.-J. Chen, 2012) could offer a path towards more accurate aggregation.

Our final research question covered the consideration of adaptability: "the state of being able to change to work or fit better" (J.-H. Cho et al., 2019). We found that very few papers consider the effects of missing data, dirty data, security events, or concept drift; all are vital elements in determining the ability of a solution to adapt to unexpected circumstances to work better. Research does often recognise the need for being able to change to fit better, as shown by the relatively large proportion that considers adaptation to other use cases. Nevertheless, there is still much to be gained in this area. It is vital that authors of research on socio-technical cybersecurity measurement solutions explicitly address the adaptability dimension in the future. Our framework of Table 3.10 helps in this regard, with its focus on proactive processes and active monitoring and detection capabilities.

We additionally analysed the ADKAR factors that were addressed in our inclusions. We found that desire was rarely considered in research. This was especially true for research focusing on the defender perspective. Additionally, we found that the real-life threat environment, as defined in Gollmann et al. (2015), is considered in less than half of our inclusions. Both of these findings offer an interesting contrast to the increasingly important role SDT and PMT play in security research (Menard et al., 2017). These theories focus heavily on (intrinsic) motivation and threat perception (Padayachee, 2012). Given the low intrinsic motivation among SMEs and their employees to improve security (Heidt et al., 2019), and the relatively large impact individual employees can have in the SME context, future research focusing on motivation and the real-life threat environment could provide an interesting avenue for making cybersecurity solutions more suitable to SMEs.

3.6.1 *Limitations and threats to validity*

We should mention at this stage that our research is not without its limitations. One potential issue is that our systematic review was not restricted to recent years, which meant that contemporary research was not as prominent in this review as it is in most other reviews. This could mean that we are overlooking certain recent developments, although 18 of our 60 inclusions were published in the past three years.

Additionally, although we believe our 60 inclusions are sufficient to help us answer our research questions, certain groupings of the inclusions resulted in relatively small sub-samples from which to draw conclusions. This could limit the generalisability of our analysis and conclusions, meaning that one could have different findings when considering different cybersecurity focus areas.

We believe in the construct validity of our systematic review methodology SYMBALS (van Haastrecht, Sarhan, Yigit Ozkan, et al., 2021), as it is based on widely-accepted methods (van de Schoot et al., 2021; Wohlin, 2014) and guidelines (Kitchenham and Charters, 2007; Liberati et al., 2009; Moher et al.,

2015). However, it is still a novel methodology that remains to be extensively tested. We feel this does not threaten the validity of our research, since SYMBALS is geared towards reproducibility and satisfies standard reporting item guidelines for systematic reviews (Moher et al., 2015).

A final mention should be made of our choice to approach the social dimension through the ADKAR change management model (Hiatt, 2006). Although the model has been applied in the cybersecurity domain (Da Veiga, 2018), it is certainly not a standard approach to use ADKAR in this setting. Nevertheless, Table 3.5 summarised the natural mapping of social cybersecurity metric concepts to the ADKAR framework and our framework presented in Table 3.10 showed how the ADKAR terms can be instinctively imported from previous research. Hence, we feel justified in using this approach.

3.7 CONCLUSION AND FUTURE RESEARCH

Businesses, and especially small- and medium-sized enterprises (SMEs), struggle to cope with the existing cyber threat landscape. Researchers have turned to cybersecurity measurement to deal with these issues, although many challenges remain, such as how to aggregate sub-metrics into higher-level metrics (J.-H. Cho et al., 2019). The challenges faced by SMEs are compounded by the dynamic nature of the cyber threat landscape, necessitating adaptable solutions. These current challenges motivated us to investigate the topics of aggregation and adaptability in this review, with a focus on SMEs.

The social side of cybersecurity deserves attention, certainly in the SME context. This is why we chose to direct our review at socio-technical cybersecurity measurement solutions. The ADKAR (Awareness, Desire, Knowledge, Ability, Reinforcement) change management model of Hiatt (2006) guided us in covering the social dimensions considered in research. To aid in the analysis of aggregation approaches, we outlined five main aggregation strategy classes in Section 3.2.3: weighted linear combinations, weighted products, weighted maxima, weighted complementary products, and Bayesian networks. We looked towards existing research to determine interesting dimensions of adaptability, such as missing or dirty data (W. Kim et al., 2003) and concept drift (Widmer and Kubat, 1996).

Based on our analysis in Sections 3.2.3 and 3.4, we found that aggregation should only be carried out if necessary, since no single aggregation strategy exists that satisfies all of the desired security properties. Notably, dependencies among metrics are often not considered. Solutions can be found in this area in Bayesian networks (Dantu and Kolan, 2005; Dantu, Kolan, and Cangussu, 2009; N. Feng et al., 2014; Sahinoglu, 2008) and analytic network process (Brožová et al., 2016; Lo and W.-J. Chen, 2012) techniques.

We used our findings as input to construct a socio-technical cybersecurity framework for SMEs. We presented our framework in Table 3.10 and visualised it in Figure 3.4. Offering a single solution for all SMEs is too simplistic. This

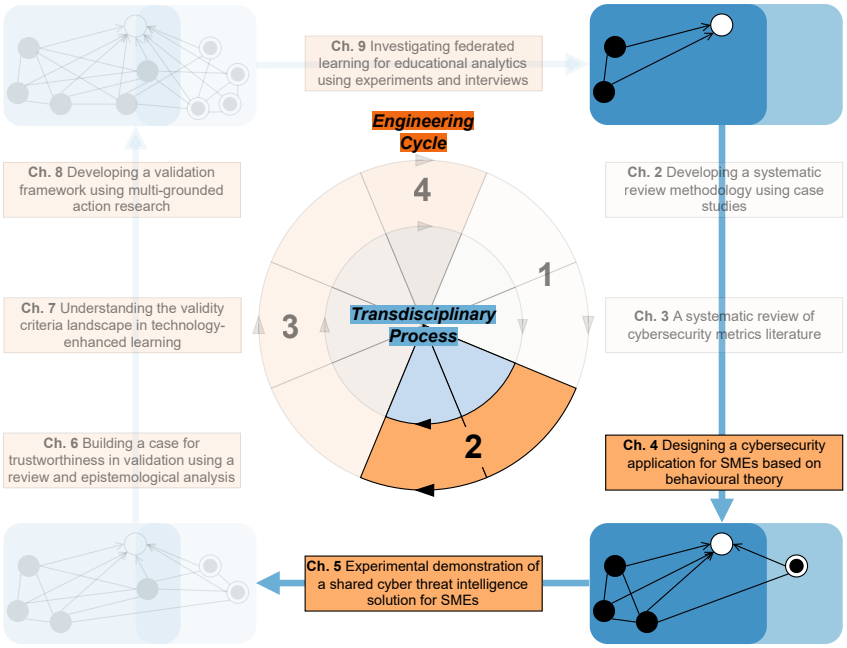
is why we divided SMEs into four categories, as suggested by the European DIGITAL SME Alliance (2020): start-ups, digitally dependent SMEs, digitally based SMEs, and digital enablers. By detailing what can be expected of each SME category, we were able to determine which cybersecurity assessment strategies were suitable in each case. For start-ups and digitally dependent SMEs, threat-based risk assessment approaches that either do not aggregate or use intuitive aggregation strategies are ideal. By focusing on the real-life threat environment (Gollmann et al., 2015), relevance and significance of the assessment task are given a central role. A simple and intuitive aggregation strategy accommodates feelings of competence and relatedness. Altogether, this ensures optimal organisation and employee motivation (Kam et al., 2020; Menard et al., 2017).

Digitally based SMEs and digital enablers are advised to use more comprehensive risk assessment approaches and maturity models. These assessment techniques should assist in working towards or proving compliance with standards and regulations. Under ideal circumstances, this will build trust in the cybersecurity posture of the SME, both internally and externally. Digital enablers are also prime candidates for using more advanced aggregation strategies such as Bayesian networks, since they often have the cybersecurity expertise and data required to make these solutions successful.

We hope that our socio-technical cybersecurity framework will provide a basis to design successful cybersecurity assessment solutions for SMEs. SMEs should not be forced to use solutions that are not suited to their situation. Especially start-ups and digitally dependent SMEs currently lack suitable cybersecurity assessment solutions, even though they are most in need of "easily understandable and practical solutions" (European DIGITAL SME Alliance, 2020). In future work, we aim to help these SMEs to become more secure. An important first step is to formulate a properly motivated, intuitive, and usable threat-based cybersecurity risk assessment approach, to offer this most vulnerable group some deserved cybersecurity respite.

Part II

TREATMENT DESIGN



4

THREAT-BASED CYBERSECURITY RISK ASSESSMENT FOR SMES

Cybersecurity incidents are commonplace nowadays, and Small- and Medium-Sized Enterprises (SMEs) are exceptionally vulnerable targets. The lack of cybersecurity resources available to SMEs implies that they are less capable of dealing with cyber-attacks. Motivation to improve cybersecurity is often low, as the prerequisite knowledge and awareness to drive motivation is generally absent at SMEs. A solution that aims to help SMEs manage their cybersecurity risks should therefore not only offer a correct assessment but should also motivate SME users. From Self-Determination Theory (SDT), we know that by promoting perceived autonomy, competence, and relatedness, people can be motivated to take action. In this chapter, we explain how a threat-based cybersecurity risk assessment approach can help to address the needs outlined in SDT. We propose such an approach for SMEs and outline the data requirements that facilitate automation. We present a practical application covering various user interfaces, showing how our threat-based cybersecurity risk assessment approach turns SME data into prioritised, actionable recommendations.

The contents of this chapter are based on: van Haastrecht, Sarhan, Shojaifar, et al. (2021). A threat-based cybersecurity risk assessment approach addressing SME needs. In Proceedings of the 16th International Conference on Availability, Reliability and Security.

4.1 INTRODUCTION

Cybersecurity incidents are commonplace nowadays and can have a devastating impact on businesses (Yigit Ozkan, van Lingen, et al., 2021). Small- and Medium-Sized Enterprises (SMEs, (European Commission, 2016)) are especially vulnerable since they have limited resources to deal with cyber-attacks (Heidt et al., 2019). Additionally, the lack of cybersecurity knowledge and awareness of SME employees causes low motivation to improve the SME cybersecurity posture (Heidt et al., 2019).

A vital first step towards managing cybersecurity risks is to assess these risks (Shameli-Sendi, Aghababaei-Barzegar, et al., 2016). Several cybersecurity risk assessment approaches tailored to SMEs exist (Mijnhardt et al., 2016; Spruit and Röling, 2014; Yigit Ozkan, Spruit, et al., 2019). From the two leading behavioural theories in the security field - Protection Motivation Theory (PMT) and Self-Determination Theory (SDT) - we know that users are most likely to take action if risk assessment solutions manage to convince the user of the risk associated with cybersecurity threats and their ability to deal with those threats (Martens et al., 2019; Menard et al., 2017; van Bavel et al., 2019). In PMT, this translates to a focus on threat- and coping appraisal (Martens et al., 2019), whereas in SDT perceived autonomy, competence, and relatedness are seen as the main drivers of motivation.

Knowing that motivation to improve cybersecurity is relatively low among SMEs (Heidt et al., 2019), it is reasonable to expect that cybersecurity risk assessment solutions for SMEs address the PMT and SDT factors. This is especially relevant for SMEs that are less digitally mature, as they are often unaware of cyber threats and require easily understandable solutions due to their limited (initial) cybersecurity knowledge (European DIGITAL SME Alliance, 2020). Sadly, most solutions are not adapted to suit SME needs (Heidt et al., 2019), with researchers insisting it is the responsibility of SMEs to take action (Benz and Chatterjee, 2020; Kaila and Nyman, 2018), rather than designing solutions that motivate SMEs (Carías, Borges, et al., 2020; Shojaifar, Fricker, and Gwerder, 2020). By not properly addressing the psychological needs identified by PMT and SDT, these solutions are much less likely to motivate SME users (Hanus and Wu, 2016).

Threat-based cybersecurity risk assessment approaches are a common tool to address the motivational issues of existing solutions. Threat-based approaches motivate threat appraisal through the incorporation of real-life threat information (Gollmann et al., 2015). Additionally, as Menard et al. (2017) recognise, any appeal for adopting cybersecurity countermeasures will be directly or indirectly based on a particular threat. Threat-based approaches offer a natural way to prioritise countermeasures, which is an important requirement in facilitating a usable solution for SMEs (Carías, Borges, et al., 2020).

It is no surprise that threat-based approaches are common in both the privacy (Deng et al., 2011; Wuyts et al., 2014) and cybersecurity (Atamli and Martin, 2014; Lippmann and Riordan, 2016; Xiong and Lagerström, 2019) fields. Threat-based cybersecurity risk assessment approaches specifically aimed at enterprises already exist (Lippmann and Riordan, 2016; B. Tucker, 2020). However, it has been well documented that approaches for enterprises in general do not map well to the SME situation (European DIGITAL SME Alliance, 2020; Heidt et al., 2019).

As a result, it is essential to discover how a threat-based cybersecurity risk assessment can be made to work for SMEs, without losing its ability to motivate users through the needs identified in PMT and SDT. This inspires the research question of this chapter:

- **RQ**: How can we create a cybersecurity risk assessment approach for SMEs that promotes user motivation?

In Section 4.2, we provide further insight into the context and motivation of this research. Section 4.3 introduces our algorithm, along with the requirements - both technically and in terms of data - for it to function properly. A practical application of our approach is outlined in Section 4.4. Section 4.5 discusses the dependencies within our solution and the privacy implications of our risk assessment approach. Finally, in Section 4.6, we conclude and propose ideas for future work.

4.2 CONTEXT AND MOTIVATION

The European Horizon 2020 project GEIGER (GEIGER Consortium, 2020) aims to help SMEs, and specifically micro-enterprises, to improve their cybersecurity posture and protect themselves against cybersecurity risks. The GEIGER project targets the smallest and least digitally mature SMEs. This group requires simple and understandable solutions, that nonetheless manage to address all areas of cybersecurity risk assessment (European DIGITAL SME Alliance, 2020). We believe a threat-centric cybersecurity risk assessment approach addresses these needs.

Cybersecurity risk assessment approaches inherently include a view on threats, due to the link between the concepts of risk and threat. At times researchers make this link explicit when employing some variant of the definition $risk = threat \times vulnerability \times consequence$ (Cox, 2008; Stergiopoulos et al., 2018). In other approaches, such as when building on the vulnerability-threat-control paradigm (C. P. Pfleeger and S. L. Pfleeger, 2012), the link is implicit, but present.

Nevertheless, we can distinguish threat-based cybersecurity risk assessment approaches - that centrally position the threat concept - from approaches that are not threat-based. In Section 4.2.1 we focus on cybersecurity risk assessment

methodologies that are aimed at SMEs and not threat-based. These approaches will often not include the real-life threat environment (Gollmann et al., 2015). Section 4.2.2 covers threat-based approaches not specifically geared towards SMEs.

4.2.1 *Cybersecurity risk assessment for SMEs*

Although SMEs are often addressed as a single group, in the cybersecurity context there are large differences among SMEs (European DIGITAL SME Alliance, 2020). This motivates a need for solutions that adapt based on the organisational characteristics of SMEs, such as the SME country or region (Sarabi et al., 2016), the SME sector (Mijnhardt et al., 2016) and the cybersecurity knowledge available in the SME (Yigit Ozkan and Spruit, 2020). The European Digital SME Alliance additionally proposes to take into account the role that an SME plays in the digital ecosystem, distinguishing four categories: digital enablers, digitally based SMEs, digitally dependent SMEs, and start-ups (European DIGITAL SME Alliance, 2020).

To attend to the needs of SMEs, certain cybersecurity risk assessment methodologies have been adapted to be suitable for smaller businesses (Alberts et al., 2005; ENISA, 2007). Maturity models are also often employed, due to their ability to provide a complete assessment while being able to adapt based on SME characteristics (Baars et al., 2016; Mijnhardt et al., 2016; Yigit Ozkan, Spruit, et al., 2019). The difficulty with all of these approaches is that they generally require a certain level of cybersecurity expertise to be present at the SME and that they assume to be dealing with a motivated user. Although these assumptions may hold for digital enablers and digitally based SMEs, this certainly cannot be expected of the digitally dependent SMEs and start-ups, who generally have little to no cybersecurity knowledge and are therefore also minimally motivated to improve their cybersecurity situation (Heidt et al., 2019).

Cybersecurity risk assessment solutions would be better suited to digitally dependent SMEs and start-ups if they could incorporate the important psychological factors outlined by PMT and SDT (Martens et al., 2019; Menard et al., 2017). Approaches explicitly incorporating behavioural theory insights are promising (Shojaifar, Fricker, and Gwerder, 2020), but contain knowledge requirements that digitally dependent SMEs and start-ups cannot fulfil. Threat-based risk assessment approaches offer interesting possibilities to assist these least digitally mature SMEs.

4.2.2 *Threat-based cybersecurity risk assessment*

Threat-based cybersecurity risk assessment approaches are not commonly applied to SMEs. That certainly does not imply, however, that these approaches

are not prominent. In privacy risk assessment, the ability to prioritise controls from a threat-based methodology is one of the reasons mentioned for preferring such an approach (Deng et al., 2011). In cybersecurity risk assessment, threat-based approaches are popular not only for their prioritisation ability (Atamli and Martin, 2014; Lippmann and Riordan, 2016; Muckin and Fitch, 2019), but also due to their ability to facilitate automation through threat catalogues (Casola et al., 2019) and publicly shared incident information (Y. Liu et al., 2015). Common risk assessment methodologies used in practice, such as STRIDE (Scandariato et al., 2015) and OCTAVE (B. Tucker, 2020), are also regularly threat-based.

The prevalence of threat-based cybersecurity risk assessment methodologies aligns with the observation that real-life threat information should be incorporated in these approaches (Gollmann et al., 2015). Threat appraisal is central in PMT and surfaces when applying SDT in the cybersecurity setting (Menard et al., 2017; Padayachee, 2012). By using insights from PMT and SDT to design appropriate nudges (Shojaifar, Fricker, and Gwerder, 2020; van Bavel et al., 2019), threat-based approaches have the potential to be highly suitable to SMEs (Y. Lee and Larsen, 2009).

We can conclude that threat-based cybersecurity risk assessment approaches can motivate SMEs to improve their cybersecurity under the right circumstances. The least digitally mature SMEs - digitally dependent SMEs and start-ups - stand to gain the most (European DIGITAL SME Alliance, 2020). Nevertheless, threat-based approaches are not commonly employed to assist SMEs. In the remainder of this chapter, we formulate a threat-based cybersecurity risk assessment approach for SMEs and argue for the motivational benefits of such an approach.

4.3 A THREAT-BASED CYBERSECURITY RISK INDICATOR

A threat-based cybersecurity risk assessment algorithm must be supported by a data model and data sources that are equally threat-centric. In this section, we describe how a threat-based view of SME cyber-systems produces a data model supporting a threat-based approach to cybersecurity risk assessment. We outline the data required to enable our approach and describe the algorithm that transforms the data into a cybersecurity risk indicator.

4.3.1 *Data model*

The impetus for an SME owner to perform a cybersecurity risk assessment is that they want to learn how to protect their SME. Figure 4.1, adapted from Casola et al. (2020), shows how this original motivation serves as one of the aspects involved in a threat-based cybersecurity risk assessment. The SME consists of assets that are valuable to the SME, such as users and devices.

The vulnerability-threat-control paradigm (C. P. Pfleeger and S. L. Pfleeger, 2012) is a general framework that can be used as a basis for our assessment approach. Within the paradigm assets can have vulnerabilities that can be exploited by threats, leading to loss or harm. Cybersecurity metrics can be used to indicate the cybersecurity risk faced by a particular asset. Cybersecurity metrics result from measuring the cybersecurity properties of an asset. The metric value should correlate to the vulnerability of the asset being measured so that it can be used in assessing risk. In this context, the risk indication given by cybersecurity metrics signifies the potential of threats to exploit vulnerabilities. To counter vulnerabilities and mitigate risk, the SME owner can enforce countermeasures, which are sometimes referred to as controls.

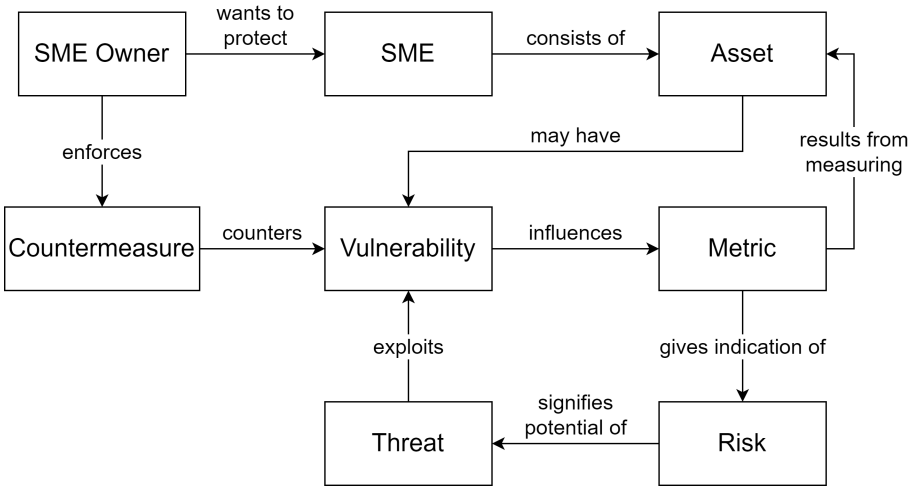


Figure 4.1: View on cyber-systems, adapted from Casola et al. (2020) to fit a threat-based cybersecurity risk assessment approach for SMEs.

Although the model in Figure 4.1 provides a clear depiction of the concepts involved in our threat-based approach, it is not detailed enough to serve as a basis for defining our algorithm data requirements. Figure 4.2, a conceptual data model, addresses this issue.

The risk profile, location, and sector elements of the enterprise entity shown in Figure 4.2 allow the algorithm to adapt based on the characteristics of the SME. Threats, metrics, and recommendations are core elements of our model. We use the term recommendation rather than countermeasure within the GEIGER solution, to distinguish the textual explanation and motivation (recommendation) - which is the element shown to the user of our application - from the action it describes (countermeasure). Both the recommendations and metrics of our solution are related to threats, which have a central position in our approach.

The metrics of our GEIGER solution measure two types of assets: users and devices. For users, we measure their knowledge and ability through interactive cybersecurity training and education. Device metrics result from the measurement of device properties by tools incorporated in the GEIGER solution. The metric values we calculate allow us to determine an indication of the cybersecurity risk faced by the SME: the GEIGER score. We can then present the user with the most relevant recommendations, where relevance is determined by the impact that the countermeasures corresponding to the recommendations have on the threats included in the GEIGER solution. The user can implement countermeasures based on the suggested recommendations, to counter vulnerabilities and mitigate risk. Implemented countermeasures lead to an improved GEIGER score.

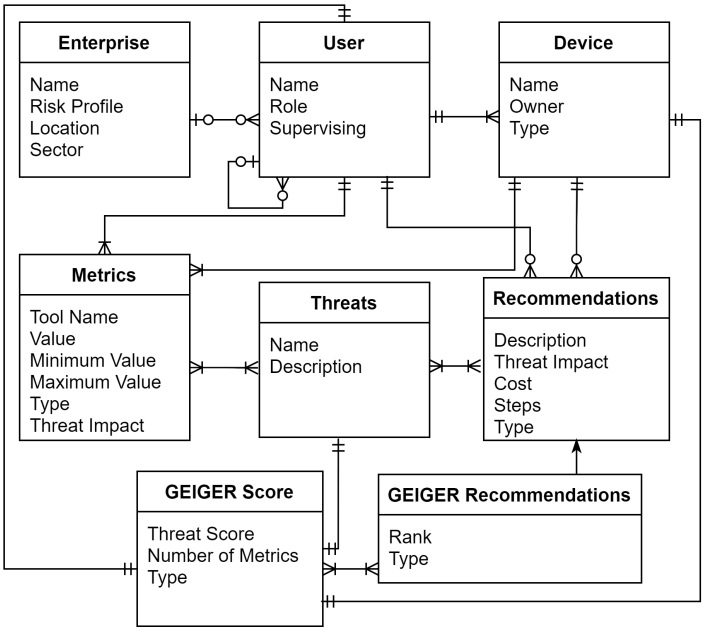


Figure 4.2: The conceptual data model underlying our threat-based cybersecurity risk assessment approach.

4.3.2 *Data requirements*

From Figure 4.2 we can derive the three main inputs required for our algorithm: metrics, threats, and recommendations. Each metric and each recommendation must relate to at least one threat.

Additionally, as discussed in Section 4.2.1, our algorithm must be able to adapt to different SME profiles. For the GEIGER project, we focus on three

specific characteristics to form the SME profile: the SME category (European DIGITAL SME Alliance, 2020), the SME country, and the SME sector. The required data then enters the system as global algorithm settings through the curator of the project, as aggregate data from Computer Emergency Response Teams (CERTs) linked to the solution, through the user entering data, or from tools that are linked to the solution. This process is depicted in Figure 4.3.

Figure 4.3 shows how users interact with the local component and how CERTs and the curator provide data to the cloud component of the solution. The local component is the application the user installs on their device. The cloud component is required to facilitate data sharing, as well as to update the algorithm based on new insights and data.

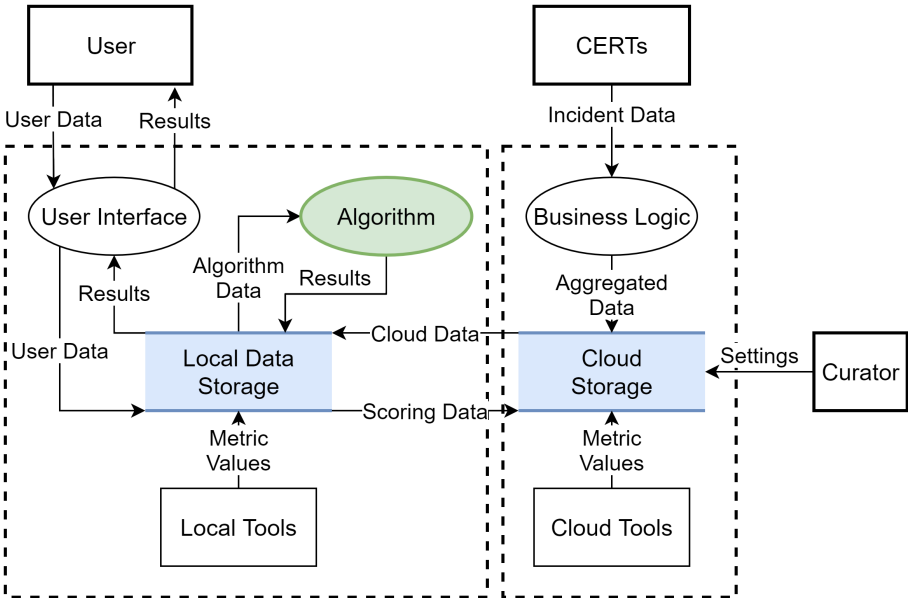


Figure 4.3: Data flow diagram showing how data from various sources flows through the system to be used in the algorithm.

To define the threats that should be considered for our SME target group, we look towards the European Union Agency for Cybersecurity (ENISA). Since 2012, ENISA publishes an annual list of top cybersecurity threats (Marinos and Sfakianakis, 2013). Through the years the list has remained remarkably unchanged, which is why it serves as an excellent basis for our threat-based approach. From the list of top threats in 2020 (ENISA, 2020), we select those threats which have been present since the first list in 2012 and are not indicated by ENISA to be part of another threat (ENISA, 2019). An exception is ransomware, which is a type of malware, but is considered to be a sufficiently significant threat to SMEs on its own to warrant inclusion.

To this set of threats, we add a threat category covering legal, third party, and supply chain threats. These three threats are a part of the general ENISA taxonomy (ENISA, 2016). They are especially relevant to our SME target group, who have a large dependency on third parties in the digital environment (European DIGITAL SME Alliance, 2020; Heidt et al., 2019). We name this category 'external environment threats', using terminology from socio-technical systems (Davis et al., 2014). This gives the following threats, in order of appearance of the ENISA top threats:

- Malware,
- Web-based threats,
- Phishing,
- Web application threats,
- Spam,
- Denial of service,
- Data breach,
- Insider threats,
- Botnets,
- Physical threats,
- Ransomware,
- External environment threats.

Figure 4.1 shows that metrics result from measuring the properties of assets within the SME. Assets in our solution are classified as employees or devices. The properties of these assets can either be measured directly, or employees of the SME can be asked to provide the necessary information on the assets. Within the GEIGER solution, we choose to (mainly) source our data from the direct measurement of asset properties by tools included within the solution. This is shown in Figure 4.3, by the data flows from local and cloud tools to their respective data storages.

Besides improving metric values, SMEs can also implement countermeasures (or controls) to counter vulnerabilities. Common countermeasures can be sourced from a variety of parties, from National Cyber Security Centres (NCSCs) and CERTs (NCSC UK, 2014; Swiss NCSC, 2021), to standards organisations (International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), 2012, 2013), to peer-reviewed research (Yigit Ozkan, van Lingen, et al., 2021). In our SME context, we should be able to argue that the countermeasures included in our solution are both

necessary and sufficient. We should not include more countermeasures than necessary, to keep our solution simple. At the same time, the countermeasures we include should be sufficient to cover all relevant areas of cybersecurity.

To address this issue we followed the following process. We first collected a large set of over 300 countermeasures from publicly available sources. We distilled this list to remove duplicates. We then mapped our list to a standard set of security countermeasure categories (International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), 2013), to see which countermeasures could be removed without losing coverage of a category. This process left a set of necessary and sufficient countermeasures, of which four examples are shown in Figure 4.4.

For a functioning threat-based cybersecurity risk assessment approach, we do not only need to define the necessary components, but we also need to determine their relationships. In our concept, both metrics and countermeasures impact threats. Furthermore, each metric and countermeasure impacts only a subset of all threats. Once tool owners and the curator of the solution have established which metrics and countermeasures relate to which threats, they must then determine impacts. To guide this process, we base ourselves on the NIST Cybersecurity Framework (Barrett, 2018), which has been used to guide cybersecurity evaluation for SMEs before (Benz and Chatterjee, 2020).

The NIST framework distinguishes five core functions: identify, protect, detect, respond, and recover. The functions can be related to various stages of a cybersecurity incident, from before the incident (identify, protect), to during the incident (detect, respond), to after the incident (recover). Since each phase is increasingly less likely to occur, the impact of countermeasures and metrics in these phases also decreases. Our approach, therefore, defines a default impact of 'high' for countermeasures and metrics relating to the identify and protect functions, 'medium' for those relating to the detect and respond function, and 'low' for those relating to the recover function.

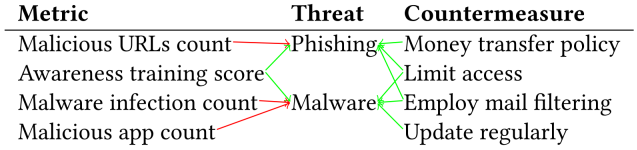


Figure 4.4: An indication of the impact of metrics and countermeasures on the common SME cybersecurity threats of phishing and malware. Green arrows indicate improving scores, whereas red arrows indicate that scores worsen.

The final piece of the puzzle, that allows us to calculate a single indicator value for an SME, is determining the relative risks associated with each threat for each SME profile. This involves making estimates of impacts and likelihoods, to calculate the common risk value: $risk = impact \times likelihood$ (Stergiopoulos et al., 2018; B. Tucker, 2020). By surveying experts as well as

security literature and reports, we can gain initial insights. However, this will not be sufficient to formulate risk estimations for each SME profile, which is an essential part of creating an adaptable approach (Baars et al., 2016).

This is why we propose to use CERT incident data to be able to create risk estimations per profile. Figure 4.3 shows how CERT incident data can be fed into our solution and aggregated, to then be used in determining threat-specific risks for each SME profile. Besides facilitating adaptability, the CERT incident data also allows us to incorporate real-life threat information into our solution. We hope this will promote perceived relatedness among SMEs.

4.3.3 *Algorithm description*

In this section, we will describe the general mathematical representation of our algorithm. An SME can be seen as a cyber-system using the definition of Refsdal et al. (2015). Similarly, each asset of the SME, such as an employee or device, can be seen as a cyber-system. This allows us to formulate an algorithm that assesses sub-systems and recursively iterates to arrive at an overall SME score.

Let S be the total set of cyber-systems of the SME, including the SME itself. Let T be the set of threats and P the set of SME profiles. Each combination of threat $t \in T$ and profile $p \in P$ has an associated relative risk $r_{pt} \in (0, 100]$.

Let M be the set of metrics. The normalised value of a metric $m \in M$ for cyber-system $s \in S$ is given by $v_{ms} \in [0, 1]$. We distinguish metrics that indicate improved security from metrics that indicate worsened security. Theoretically, a single metric may even relate positively to security for one threat, but negatively for another. Hence, we define the Boolean indicator δ_{mt}, which equals 1 when a metric $m \in M$ relates positively to the relative risk associated with threat $t \in T$.

We further define the impact of metric $m \in M$ on threat $t \in T$ as i_{mt}. Recall that this impact may either be low, medium or high. We map these categories to values of 0.1, 0.5, and 1.0, respectively. To be able to keep track of which metrics have been calculated, we define the Boolean variable λ_{ms}, which equals 1 if metric $m \in M$ has been calculated for cyber-system $s \in S$.

We let C be the set of countermeasures. The variable i_{ct} has an identical definition as in the metric case. The Boolean variable λ_{cs} is now used to indicate whether a countermeasure $c \in C$ has been implemented for cyber-system $s \in S$. Since we only allow for countermeasures to be implemented or not implemented, without assigning a specific value, we have no analogue for the variable v_{ms} specifying the metric value. Similarly, since countermeasures always relate positively to security, there is no analogue to the δ_{mt} variable.

All of our defined variables allow us to calculate the indicator value I_{spt} specific to threat $t \in T$, for a cyber-system $s \in S$, which is (part of) an SME with profile $p \in P$:

$$I_{spt} = 50 + 50 \cdot \frac{\sum_{m \in M} \delta_{mt} \cdot \lambda_{ms} \cdot i_{mt} \cdot v_{ms}}{\sum_{m \in M} \delta_{mt} \cdot \lambda_{ms} \cdot i_{mt}} - 25 \cdot \left(\frac{\sum_{m \in M} (1 - \delta_{mt}) \cdot \lambda_{ms} \cdot i_{mt} \cdot v_{ms}}{\sum_{m \in M} (1 - \delta_{mt}) \cdot \lambda_{ms} \cdot i_{mt}} + \frac{\sum_{c \in C} \lambda_{cs} \cdot i_{ct}}{\sum_{c \in C} i_{ct}} \right). \tag{4.1}$$

Equation 4.1 ensures the indicator value I_{spt} ranges from 0 to 100 and initially takes a value of 50. Note that our current assumption is that countermeasures always apply to all cyber-systems under consideration. However, if necessary, the algorithm could easily be extended with an additional Boolean variable to permit variation in this dimension.

Some of the divisors of Equation 4.1 equal 0 when no values have been calculated. In this scenario, we set the value of the relevant fraction to 0. The total indicator score over all threats, again ranging between 0 and 100, is given by:

$$I_{sp} = \frac{\sum_{t \in T} I_{spt} \cdot r_{pt}}{\sum_{t \in T} r_{pt}}. \tag{4.2}$$

In essence, Equation 4.2 could be used to calculate the indicator value for the complete SME, if the system $s \in S$ considered is the SME itself. However, in practice, there are privacy constraints to sharing all data within the full company. Some of this data, especially the security information related to employees, can be sensitive. So, we need to formulate a process to arrive at an indicator value representing the entire SME, without needing to share all data items.

To solve this issue we recognise that SMEs, like any enterprise, are generally hierarchically structured. The owner of the SME is positioned at the top of the hierarchy and supervises one or more employees. These employees, in turn, may supervise further employees. By incorporating this supervision structure in our scoring mechanism, we can ensure that a minimal amount of data is shared, while still arriving at an indicator value that accurately represents the complete SME.

Within our approach, we distinguish two types of scores: user scores and device scores. User scores relate to the knowledge and ability of an employee within the SME, whereas device scores relate to the security properties of the device. Each employee $e \in S$ that has installed the GEIGER application on a device they own will therefore have at least two scores: their user score and the score of the device they own. An employee may own multiple devices and can therefore have more than two associated scores.

An employee may not wish to share their user score per threat with their supervisor, due to the sensitive nature of this information. This is why we propose to only share aggregated data. Let n_s be the total number of metrics

calculated to arrive at the indicator value for cyber-system $s \in S$. Based on our earlier definitions, we have:

$$n_s = \sum_{m \in M} \lambda_{ms}.$$

We define the set of employees $E \subset S$, to help in addressing supervision. We then define $S_e \subseteq S$ to be the set of cyber-systems belonging to employee $e \in E$. This set corresponds to the employee themselves and the devices they own. Let $E_e \subset E$ be the set of employees supervised by employee $e \in E$. We then define the aggregate score of employee $e \in E$ as:

$$I_{ep}^{agg} = \frac{\sum_{s \in S_e} I_{sp} \cdot n_s + \sum_{\hat{e} \in E_e} I_{\hat{e}p}^{agg} \cdot n_{\hat{e}}^{agg}}{\sum_{s \in S_e} n_s + \sum_{\hat{e} \in E_e} n_{\hat{e}}^{agg}}, \tag{4.3}$$

where:

$$n_e^{agg} = \sum_{s \in S_e} n_s + \sum_{\hat{e} \in E_e} n_{\hat{e}}^{agg}. \tag{4.4}$$

The recursive nature of Equation 4.3 and Equation 4.4 allow us to iteratively calculate aggregate scores until we reach the aggregate score of the SME owner. The aggregate score of the SME owner represents all of the information available for scoring, and therefore accurately represents the cybersecurity posture of the SME. Since only aggregate data is shared, the scoring procedure preserves privacy while still managing to achieve an accurate score. Table 4.1 provides an overview of all of the variables discussed in this section.

The formulation of our algorithm allows us to determine the place our threat-based cybersecurity risk assessment approach takes within the information security risk assessment (ISRA) taxonomy of Shameli-Sendi, Aghababaei-Barzegar, et al. (2016). Our approach is quantitative and asset-driven. Additionally, assets are evaluated independently of each other and risk assessment scores are propagated through recursive formulas. Furthermore, we do not assign a monetary value to assets.

Based on the ISRA taxonomy, our approach is similar to other risk assessment approaches (Alpcan and Bambos, 2009; Ben Mahmoud et al., 2011; Schmidt and Albayrak, 2010). However, none of these methodologies uses threat-based techniques, nor do they use the hierarchical structure we propose to use for SMEs. We can conclude that although our approach follows established guidelines for formulating a cybersecurity risk assessment methodology, it has unique elements. These elements are included to make our approach suitable for SMEs. The following section provides further explanation on how our algorithm results are translated into visual representations to effectively nudge SME users.

Table 4.1: The variables used within the algorithm.

VARIABLE	DEFINITION
S	The set of all cyber-systems within the SME.
S_e	The set of cyber-systems belonging to employee $e \in E$, $S_e \subseteq S$.
E	The set of all employees within the SME, $E \subset S$.
E_e	The set of employees supervised by employee $e \in E$, $E_e \subset E$.
T	The set of all threats.
P	The set of all SME profiles.
M	The set of all metrics.
C	The set of all countermeasures.
r_{pt}	Relative risk of threat $t \in T$, for profile $p \in P$.
v_{ms}	Normalised value of metric $m \in M$, for cyber-system $s \in S$.
i_{mt}	Impact of metric $m \in M$ on threat $t \in T$.
i_{ct}	Impact of countermeasure $c \in C$ on threat $t \in T$.
λ_{ms}	Boolean variable equalling 1 when metric $m \in M$ has been calculated for cyber-system $s \in S$.
λ_{cs}	Boolean variable equalling 1 when countermeasure $c \in C$ is implemented for cyber-system $s \in S$.
δ_{mt}	Boolean variable equalling 1 when metric $m \in M$ relates positively to the risk of threat $t \in T$.
I_{spt}	Threat-specific cybersecurity risk indicator for cyber-system $s \in S$.
I_{sp}	Cybersecurity risk indicator for cyber-system $s \in S$.
i_{ep}^{agg}	Aggregate cybersecurity risk indicator for employee $e \in E$.
n_e^{agg}	Total number of metrics calculated to arrive at i_{ep}^{agg}.

4.4 EXEMPLAR OF PRACTICAL APPLICATION

Self-Determination Theory (SDT) is a theoretical framework used in the study of motivational dynamics and individual behaviours (Deci and Ryan, 1985; Ryan and Deci, 2000). SDT distinguishes intrinsic and extrinsic types of motivation and explains people's psychology of being self-determined to adopt behaviour and persist in an activity. SDT elaborates three fundamental psychological needs – autonomy, competence, and relatedness – and assumes that their satisfaction leads to self-motivation, engagement, and positive outcomes (Vallerand, 1997).

- **Autonomy**: A desire to engage in activities with willingness and a freedom of choice,
- **Competence**: A desire to interact effectively with the environment for developing wanted outcomes and preventing undesired events,
- **Relatedness**: A sense of belongingness and connectedness to others or a social environment.

SDT is applied in cybersecurity (Menard et al., 2017) and security solution design (Shojaifar and Fricker, 2020; Shojaifar, Fricker, and Gwerder, 2020) to explain the relationships between design features and user motivation in cybersecurity. The basic psychological needs are reliable mediators to study how security tool features support user need satisfaction and consequent tool adoption. This section presents the main GEIGER toolbox interfaces and

outlines how the toolbox features operationalised SDT constructs (autonomy, competence, and relatedness) to encourage users to adopt GEIGER for protecting their companies.

4.4.1 *Main interface*

The structure of the main screen depicted in Figure 4.5 follows the approach that the most important elements are displayed on top. If a risk scan has already been carried out, the first thing the user sees is their aggregated score, which is displayed in green (low), yellow (medium), orange (high), or red (very high), depending on the level of the risk. This gives a first impression of the overall risk potential and should trigger the need to act depending on the threat situation.

The score is shown noticeably large because it is an aggregation of the user scores and the device scores across all threats. Depending on the role of the user, the labelling of the score adapts to convey whether the score represents the whole company or just one person with its employees. The aggregated score and its colour support the user's familiarity with the overall potential risks in the company and motivate the user for a desirable practice.

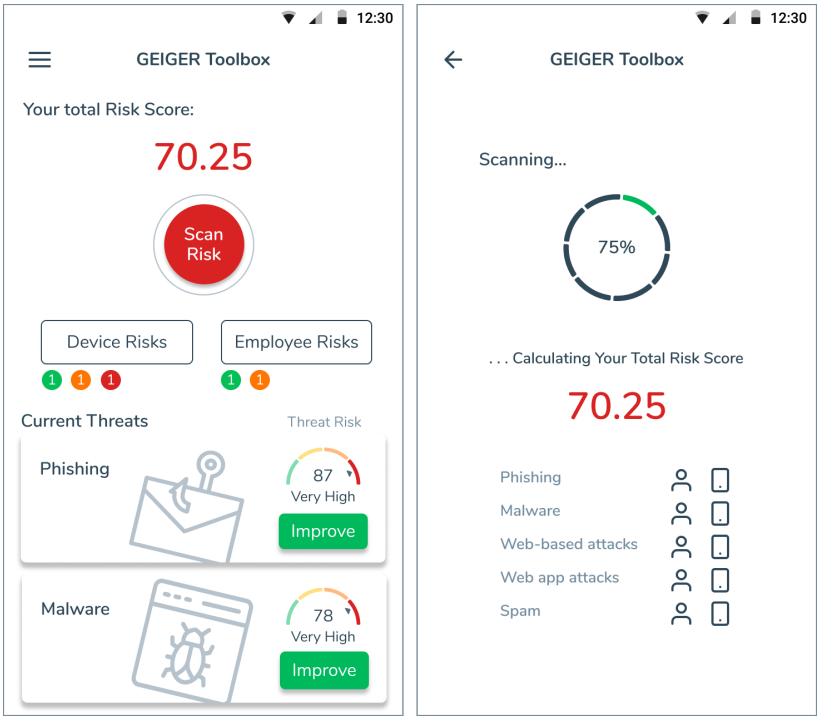


Figure 4.5: Main interface (left) and score calculation process (right).

By pressing the scan risk button, the calculation of the latest risk score is initiated. An intermediate screen shows that the app is working in the background and how far advanced the calculation process is. Furthermore, during this waiting period, the user should be shown how their aggregated score is achieved, as well as those of the employees they supervise. As soon as the calculation process has finished, the main screen will be shown again with the current aggregated score of the user as well as all threats with their current scores.

Threats with higher risk scores are shown first. Each threat is shown as a so-called card with the threat name, a threat visualisation, a threat score, and a button that leads to the recommendations for a threat. The button 'improve' is coloured green, which contrasts with the colours of high-risk scores to convey a positive action.

To get a quick overview of the situation of other devices or employees, the coloured dots below the buttons show how many devices or employees have been classified with which risk level (left image of Figure 4.5).

4.4.2 *Device and employee risk*

Using the buttons 'device risks' and 'employee risks' of Figure 4.5, the user can either navigate to a list with all their devices or to a list with all their employees. Here, the aggregated scores over all threats are displayed for each device or employee (Figure 4.6). The employee and device lists help the user to better handle security measures in the company. Moreover, the prioritised list of visualised threats and texts and the available tailored recommendations support user competence and autonomy.

In general, as soon as a scan is carried out, the scores of the devices are no longer up to date. This is depicted in the device risk screen of Figure 4.6. The device is marked and the user is prompted to open the app on the device and perform a scan.

In the case of employees, when the supervisor scans, they receive a request to allow or deny sharing their scores with their supervisor. For this reason, either the score is displayed on the employee screen if permission has been granted, or the score is displayed as pending or rejected (right image of Figure 4.6). Information sharing in GEIGER is based on users' permission. A user may choose to allow or deny sharing their information with the supervisor, stimulating perceived autonomy.

4.4.3 *Recommendations*

Using a tab, the user can switch between user- and device-specific recommendations and sees the respective score directly on the tab (left image of Figure

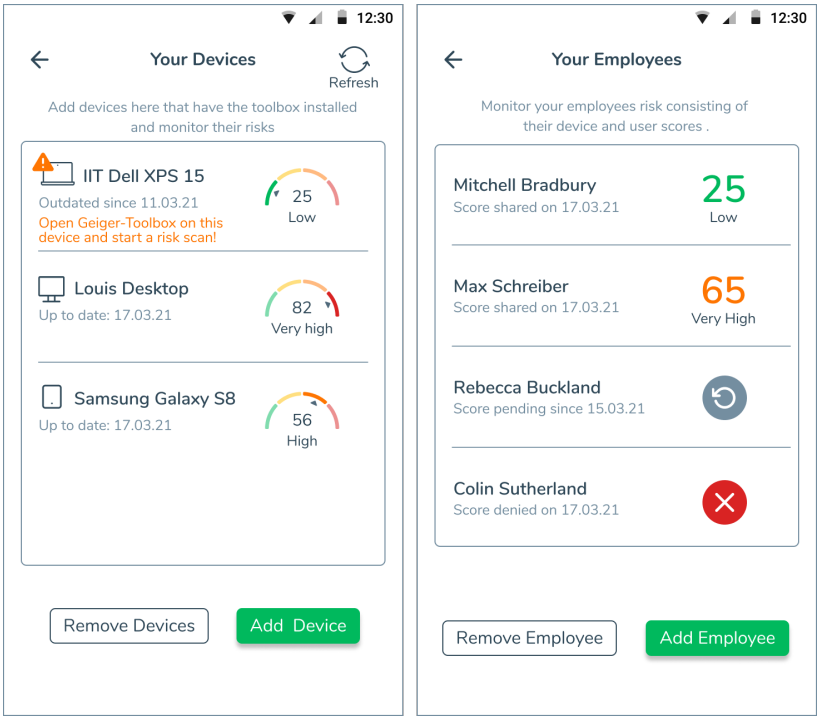


Figure 4.6: Interfaces of all devices (left) and all employees (right) with respective risk scores.

4.7). Depending on the tab, the user is shown either their name or that of their active device.

Since the target group may still be unfamiliar with threat terminology or with the concept of user and device scores, they are given the opportunity to obtain additional information. Figure 4.7 shows how this information can be accessed, for example, via a button labelled 'About Phishing' or 'About User Score.' To prevent flooding the user with information, the respective input is presented in the form of several small blocks and with corresponding illustrations.

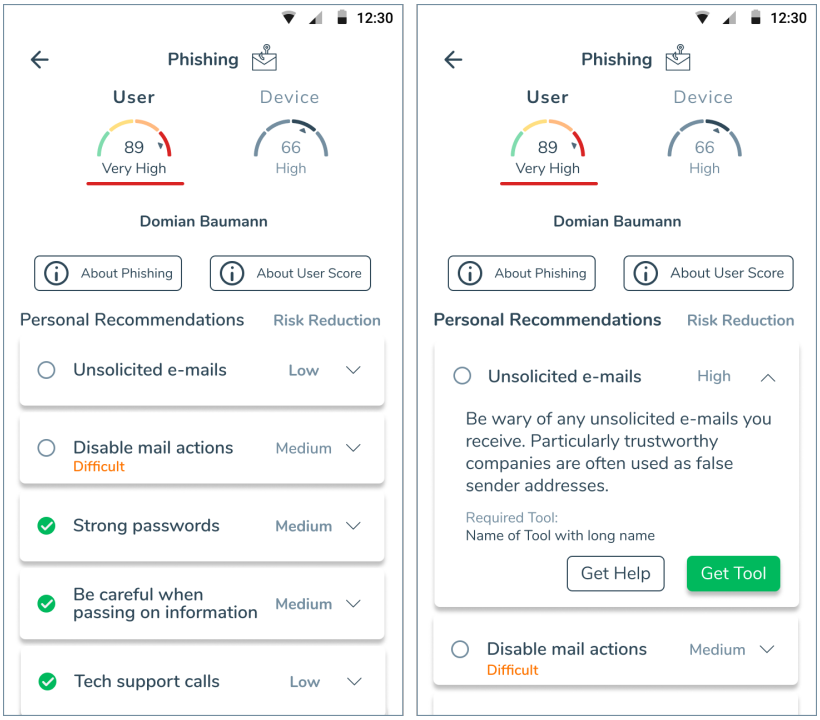


Figure 4.7: Interfaces of user-specific recommendations for phishing.

The recommendations with the highest impact on risk reduction are displayed, given that they correspond to the knowledge level of the user and are yet to be implemented. Recommendations that have been implemented are marked with a green tick. Each recommendation is categorised with a risk reduction impact of low, medium, or high.

The recommendations can contain learning content so that the user is more likely to recognise dangers and improve their behaviour in the long term. There are also recommendations in which the user must implement a precautionary measure, guided by step-by-step instructions. The user can

implement some recommendations directly with the help of the app, while others require additional tools that take on more complex tasks.

Recommendations that could be too demanding are marked as 'Difficult,' whereby the user is asked to contact a security defender if necessary. In any case, the user can use the 'Get Help' button to access a list of security defenders to receive more personal support. The recommendation support embedded in GEIGER helps to promote perceived autonomy and competence. By enabling contact with a trusted advisor, in the form of a security defender, we hope to stimulate perceived relatedness and competence among users.

The GEIGER features are designed to provide information and familiarity with different types of potential security threats and improve user experience. Various colours and scores support users' appraisal of the risks, and in turn, support extrinsic motivation to enact security measures (Padayachee, 2012). Consistent with SDT's three basic psychological needs, GEIGER features are designed to facilitate daily self-determined cybersecurity improvement.

4.5 DISCUSSION AND LIMITATIONS

The GEIGER indicator relies on several threat-related metrics collected by different GEIGER tools to provide relevant insight into the risk level of an SME, including its devices and employees. This module is part of the GEIGER ecosystem composed of scanning tools (for threat detection), education tools (for training) and components integrating data coming from different CERTs.

The confidence in the GEIGER indicator depends on the completeness of the collected data. In other words, the more data that is available and recent, the more accurate the GEIGER indicator is. Ideally, the uncertainty associated with a lack of data would be quantified and communicated to the user. Although this is currently not part of the GEIGER user interface, it could prove to be a valuable addition.

The GEIGER solution is composed of several interdependent components. The accuracy of the GEIGER indicator may come at a cost; the cost of complexity. We should take care to translate this underlying complexity into a simple and clear message to the user, which is what we aim to achieve with the user interface outlined in Section 4.4.

An important facet in harbouring user trust is adequately addressing confidentiality concerns (Shojaifar and Fricker, 2020). The GEIGER indicator is computed for each employee and no sharing - to the employees' supervisor or the GEIGER cloud - is allowed before the consent of this employee. The GEIGER indicator is GDPR-compliant by respecting user preferences regarding data privacy.

Yet, we wish to go further than just compliance. Since the accuracy of the GEIGER indicator is largely determined by the amount of data underlying its value, it will be necessary to create a comfortable environment for the user to provide consent to information sharing (Shojaifar and Fricker, 2020).

However, we recognise that it will be challenging to find the right balance between pushing users to share data and providing a comfortable setting, as these are somewhat conflicting goals.

The GEIGER indicator is still in its prototype release. More validation with end-user SMEs is planned in the coming months to refine its scope and improve its reliability in terms of the suggested recommendations to protect SMEs from the most impactful cyber-threats.

4.6 CONCLUSION AND FUTURE WORK

Less digitally mature Small- and Medium-Sized Enterprises (SMEs) are perhaps the most vulnerable to cybersecurity threats of all organisations. These SMEs often lack the cybersecurity knowledge, awareness, and resources to deal with cyber-attacks. Perhaps even more worryingly, their limited connection to the cybersecurity topic often causes a low motivation to improve their cybersecurity posture. This is why we set out to answer the question: How can we create a cybersecurity risk assessment approach for SMEs that promotes user motivation?

Any appeal for adopting cybersecurity countermeasures is, directly or indirectly, motivated by a particular threat. Unsurprisingly, threat-based cybersecurity risk assessment methodologies are a popular tool. Besides having a natural ability to promote threat appraisal, an important concept in behavioural theories such as Protection Motivation Theory (PMT) and Self-Determination Theory (SDT), threat-based approaches facilitate automation and prioritisation.

Nevertheless, threat-based cybersecurity risk assessment approaches are not commonly used to assist SMEs. We introduced a threat-based cybersecurity risk indicator specifically aimed at SMEs and discussed the data requirements to make the algorithm behind such an indicator work. After outlining the details of our algorithm, we covered a practical application of our approach, delineating how different user interface screens satisfied the three SDT needs: autonomy, competence, and relatedness.

Our work shows that it is feasible to create a cybersecurity risk assessment approach for SMEs that promotes user motivation. We strongly believe that threats should play a central role in any such solution.

We recognise that challenges remain and that more validation of our approach is necessary. In future work, we plan to refine our algorithm through the incorporation of extensive user feedback. Additionally, we intend to further investigate threat prioritisation and the possibilities of incorporating privacy-preserving ideas in our algorithm. We hope that the new insights we gain will bring the most vulnerable SMEs another step closer to security.

5

A SHARED CYBER THREAT INTELLIGENCE SOLUTION FOR SMES

The contents of this chapter are based on: van Haastrecht, Golpur, et al. (2021). A Shared Cyber Threat Intelligence Solution for SMEs. Electronics.

Small- and medium-sized enterprises (SMEs) frequently experience cyberattacks, but often do not have the means to counter these attacks. Therefore, cybersecurity researchers and practitioners need to aid SMEs in their defence against cyber threats. Research has shown that SMEs require solutions that are automated and adapted to their context. In recent years, we have seen a surge in initiatives to share cyber threat intelligence (CTI) to improve collective cybersecurity resilience. Shared CTI has the potential to answer the SME call for automated and adaptable solutions. Sadly, as we demonstrate in this chapter, current shared intelligence approaches scarcely address SME needs. We must investigate how shared CTI can be used to improve SME cybersecurity resilience. In this chapter, we tackle this challenge by using a systematic review to discover current state-of-the-art approaches to utilising shared CTI. We find that threat intelligence sharing platforms such as MISP have the potential to address SME needs, provided that the shared intelligence is turned into actionable insights. Based on this observation, we developed a prototype application that processes MISP data automatically, prioritises cybersecurity threats for SMEs, and provides SMEs with actionable recommendations tailored to their context. Our application will increase SME cybersecurity awareness and resilience, which will enable them to thwart cyberattacks in future.

5.1 INTRODUCTION

The cybersecurity threat landscape is diverse and dynamic, as witnessed by several recent supply chain attacks with worldwide impact (Browning, 2021; Lazarovitz, 2021). Attack sophistication is increasing (Skopik et al., 2016) and it is now widely accepted that even nation-states are actively involved in the most advanced and persistent threats (Lemay et al., 2018). Unsurprisingly, the trend of increased complexity in attacks is expected to continue in the future (Lella et al., 2021).

These observations stand in stark contrast to the situation of small- and medium-sized enterprises (SMEs), who lack the knowledge and resources to appropriately address any cybersecurity threats (Heidt et al., 2019); never mind advanced threats. SMEs require the help of their external environment to deal with cybersecurity attacks since they do not have internally available expertise (van Haastrecht, Yigit Ozkan, et al., 2021).

In this sense, the maxim "a problem shared is a problem halved" is fitting in the SME context. It is this maxim that is the driving force behind information sharing in the cybersecurity community (Skopik et al., 2016). Sharing cybersecurity intelligence has long been recognised as a key ingredient in raising our collective cybersecurity resilience. Yet, until recently, efforts in this area were fragmented and unsuccessful (Kampanakis, 2014), with many feeling the advantages to sharing data were outweighed by the disadvantages (Albakri et al., 2018; Ring, 2014).

This changed with the introduction of standardised cybersecurity intelligence taxonomies (Barnum, 2012; Burger et al., 2014; Connolly et al., 2012) and intelligence sharing platforms (Sauerwein et al., 2017; Wagner et al., 2016). Especially the sharing of threat (Johnson et al., 2016; Mavroeidis and Bromander, 2017; Qamar et al., 2017) and incident (Baesso Moreira et al., 2018) information gained acceptance and popularity.

Privacy concerns still remain regarding the sharing of cybersecurity intelligence (Shojaifar and Fricker, 2020; Zibak and Simpson, 2019). However, the focus has now shifted to finding solutions rather than simply detailing problems (Azad et al., 2021; de Fuentes et al., 2017; Ezhei and Tork Ladani, 2017). Exploiting the properties of blockchain for privacy preservation is an example of a novel and promising approach (Brotsis et al., 2019; Purohit et al., 2020).

Recently, the use of advanced data analytics (Husák, Komárková, et al., 2019; N. Sun et al., 2019) and machine learning (Sarker, Furhad, et al., 2021; Sarker, Kayes, et al., 2020) techniques to extract further insights from shared intelligence has spurred on optimism regarding the future of cybersecurity information sharing. Nevertheless, the literature remains eerily silent regarding the use of shared incident data to support SMEs; a group in dire need of help from their external environment.

SMEs have their own concerns regarding information sharing (Shojaifar and Fricker, 2020), and certainly require different treatments and solutions than other enterprise types (Yigit Ozkan, Spruit, et al., 2019). This is perhaps most true for the least digitally mature SME categories: *start-ups* and *digitally dependent SMEs*. Along with the more mature *digitally based SMEs* and *digital enablers*, the European DIGITAL SME Alliance (European DIGITAL SME Alliance, 2020) distinguishes these SME categories to emphasise that SMEs are not one homogeneous group, but rather a diverse set of businesses, with diverse needs.

SMEs require distinctly different solutions than other enterprises due to their lack of internally available cybersecurity knowledge and resources. Additionally, any solution looking to aid SMEs should recognise the heterogeneity within this group of enterprises. Based on what we know of current trends in cybersecurity intelligence sharing literature, it is therefore unlikely that any of the prevailing approaches to utilising shared incident data are suitable for SMEs. Nevertheless, it can be expected that current approaches contain building blocks for useful SME approaches, especially due to the automatic nature of today's machine learning techniques.

Finding out how we can use shared cybersecurity information to aid SMEs is our main focus in this chapter. Hence, we ask:

- **RQ**: How can shared incident information be utilised to help improve SME cybersecurity?

We will answer our research question by first systematically reviewing current approaches to utilising shared incident data in Section 5.2. Here we will also provide a detailed analysis of the difficulties of using the VERIS Community Database (VCDB) (*The VERIS Community Database* 2021) in the SME context. These efforts will provide insight into what adaptations to current approaches are necessary to yield a useful solution for SMEs.

We then describe our proposed solution using the Malware Information Sharing Platform (MISP) (Wagner et al., 2016) in Section 5.3, covering the input (5.3.1), process (5.3.2), and output (5.3.3). In Section 5.3.4, we provide a practical example of how our application helps SMEs, demonstrating the potential impact of our solution. Finally, we discuss our findings in Section 5.4 and conclude in Section 5.5.

5.2 LITERATURE REVIEW

Before proposing our methodology, we should investigate current approaches to utilising shared cybersecurity threat intelligence. We conducted this investigation via a systematic literature review using the SYMBALS (van Haastrecht, Sarhan, Yigit Ozkan, et al., 2021) methodology. We searched the Scopus database for the keywords presented in Table 5.1, where we restricted our search to conference and journal articles and English-language documents.

Table 5.1: Keywords and accompanying synonyms used in our search of the Scopus database.

KEYWORD	SYNONYMS
cybersecurity	cyber security, information security
threat	event, attack, incident
sharing	share

Additionally, we focused on research published since 2016. In 2016, the Malware Information Sharing Platform (MISP) was introduced (Wagner et al., 2016). MISP is one of the most widely used threat sharing platforms, along with the Trusted Automated eXchange of Indicator Information (TAXII) (Connolly et al., 2012). Both MISP and TAXII facilitate information exchange using the Structured Threat Information eXpression (STIX) language (Barnum, 2012), the de-facto standard format for exchanging threat intelligence.

The choice to focus our review on the period since 2016 is no coincidence. Since the introduction of MISP, the subject matter of shared threat intelligence research has shifted. Whereas earlier research explored information sharing options (Kampanakis, 2014; Steinberger et al., 2015) and outlined the barriers to sharing (Ring, 2014), research since 2016 has largely centred around how we can use shared intelligence.

Our database search yielded 546 results, of which 47 inclusions remained after applying the filtering steps of SYMBALS. The most common reason for exclusion was that a paper did not cover our topic of interest: the utilisation of shared threat intelligence. This is not surprising, as the keywords we employed do not provide a guarantee of papers in our focus area.

We then proceeded to extract relevant data from our inclusions. One dimension we considered was the suitable organisation type for an approach. The European DIGITAL SME Alliance outlines four SME categories: start-ups, digitally dependent SMEs, digitally based SMEs, and digital enablers (European DIGITAL SME Alliance, 2020). The cybersecurity maturity of these SME categories progresses from the least mature start-ups to the most mature digital enablers (van Haastrecht, Yigit Ozkan, et al., 2021).

Where start-ups are only beginning to realise the importance of cybersecurity, we can expect digital enablers to have embedded, automated cybersecurity processes (van Haastrecht, Yigit Ozkan, et al., 2021). Nevertheless, even digital enablers are unlikely to have the capacity to run a Security Operations Centre (SOC) which can monitor and analyse continuously gathered internal security intelligence. This is why we included a 'large enterprises' category to collect any methods unsuited to any SME category. The first column of Table 5.2 depicts our considered enterprise categories.

Ramsdale et al. (2020) offer a concise classification of cyber threat intelligence (CTI) sources. They divide sources into internally sourced intelligence, externally sourced intelligence, and open-source intelligence. Internally

Table 5.2: The type of cyber threat intelligence used in each of our 47 inclusions, along with the minimum SME category maturity required to implement the proposed methodology.

CATEGORY	EXTERNAL INTELLIGENCE	OPEN-SOURCE INTELLIGENCE	INTERNAL INTELLIGENCE
Start-ups	Vakilinia et al. (2018)	Badsha et al. (2019)	
Digitally dependent		S. He, G. M. Lee, et al. (2016)	
Digitally based	Tanrıverdi and Tekerek (2019) Riesco, Larriva-Novo, et al. (2020)	Qamar et al. (2017) Faiella et al. (2021) J. Zhao et al. (2020) Ural et al. (2021)	Brotsis et al. (2019) Best et al. (2017)
Digital enablers	Y. Zhao et al. (2017) Gonzalez-Granadillo et al. (2019) Ansari et al. (2020)	Husari et al. (2018) W. Yang and Lam (2020) Koloveas et al. (2021) Khramtsova et al. (2020) Mutemwa et al. (2017)	Purohit et al. (2020) H. Zhao and Silverajan (2020) Lin et al. (2019) Serketzis et al. (2019) Mohasseb et al. (2020) Y. Sun et al. (2020) Husák, Bartoš, et al. (2021) Jeng et al. (2019) Husák, Bajtoš, et al. (2020) Huang et al. (2020) Riesco and Villagrá (2019)
Large enterprises	E. Kim et al. (2018) S. He, Fu, et al. (2020) Schlette et al. (2021) Schaberreiter et al. (2019) Settanni et al. (2017) Manfredi et al. (2021)	Mtsweni et al. (2016) J. Yang et al. (2020)	Takahashi and Miyamoto (2016) Kure and Islam (2019) Graf and King (2018) S. Brown et al. (2019) Leszczyna and Wróbel (2019) Badri et al. (2016) McKeever et al. (2020) Abe et al. (2018) Leszczyna, Wallis, et al. (2019)

sourced intelligence relates to data on events occurring within an organisation's IT infrastructure. External intelligence comes from structured threat intelligence feeds, such as those sourced from the TAXII and MISP platforms. Finally, open-source intelligence is defined as intelligence from publicly available sources such as news feeds and social media. We choose to not employ the commonly used abbreviation of open-source intelligence OSINT, as OSINT is more broadly associated with the methodology of collecting threat intelligence from publicly available sources.

Table 5.2 categorises our inclusions based on the suitability of their approach to different enterprise types and the type of intelligence source they build on. We should note that the enterprise categories of Table 5.2 are ordered by cybersecurity maturity. This means that if start-ups can use a particular approach, digitally dependent SMEs will automatically also be able to use that approach. Similarly, if an approach is classed as being suitable for digitally based SMEs, it is not suitable for the less digitally mature start-ups and digitally dependent SMEs.

The first thing to notice about Table 5.2 is that very few of our inclusions specify shared CTI solutions suitable for start-ups and digitally dependent SMEs. We cannot expect these SMEs to collect and analyse internal intelligence, which explains why none of the internal intelligence approaches is suited to start-ups and digitally dependent SMEs. Internal intelligence approaches often require an internal security expert or even a SOC, which make them difficult to implement even for digitally based SMEs and digital enablers.

Open-source intelligence methodologies often suffer from their open-ended nature, making them less actionable for SMEs. The collected data is often

unstructured text and will generally only serve to inform the user, rather than assist them in concrete tasks. The two open-source approaches that are suited to less digitally mature SMEs have a very specific goal. In the first, the authors create a spam filter based on open-source spam data, which can then be used by organisations to prevent spam from reaching employee inboxes (Badsha et al., 2019). The second approach also uses publicly available spam data, but this time it is connected to organisation IPs and used as a tool to confront companies with their security level (S. He, G. M. Lee, et al., 2016).

Although the mentioned open-source intelligence sharing methods have their merits for start-ups and digitally dependent SMEs, they only scratch the surface of what can be done to help SMEs. Structured external intelligence could be an outcome here, but, as Table 5.2 shows, most research is geared towards large enterprises. All of the external intelligence approaches for large enterprises use STIX as their data sharing format, and most use TAXII as the sharing platform. The benefit of STIX is that it is flexible and therefore facilitates many different indicators of compromise (IoCs). However, most research proposes methodologies whereby the STIX data is shared without much processing. This means the shared data retains much of STIX's complexity, and it is left to analysts at an organisation to interpret this data. SMEs simply do not have the resources for such activities.

The external intelligence approaches suited to SMEs still regularly employ STIX. However, they no longer use TAXII as a sharing platform, preferring less common platforms or a custom sharing platform. Approaches that apply a more extensive filtering process to provide organisations with concise insights are most suited to the least digitally mature SMEs. By comparing shared data to blacklists (Tanrıverdi and Tekerek, 2019) or using the shared intelligence to advise on suitable production rules (Riesco, Larriva-Novo, et al., 2020), digitally based SMEs are aided in their detection process. However, detection is still a step too far for start-ups and digitally dependent SMEs, who are often still in the process of understanding their assets and attack surface (van Haastrecht, Yigit Ozkan, et al., 2021).

The external intelligence approach suited to start-ups uses a feed of passwords identified in breaches to inform users of susceptible passwords (Vakilinia et al., 2018). As with the open-source intelligence approaches, it is the focused nature and clear aim of this approach that makes it accessible to all types of SMEs. The question remains whether we can go beyond these specific implementations while maintaining usability for the least digitally mature SMEs. Such solutions currently do not exist and would be immensely beneficial to SMEs.

We certainly believe it is possible to create such solutions. It is clear from our systematic review results that the solution lies in the use of structured external threat intelligence, preferably conforming to the STIX standard, which is sufficiently processed and filtered to yield actionable insights for SMEs. Section 5.3 explains our solution.

Before diving into our solution, it is worth investigating whether a similar approach using open-source intelligence would also be feasible. We noted earlier that one of the main issues with open-source intelligence for SMEs is its unstructured nature. However, structured open-source intelligence sources do exist. The VERIS Community Database (VCDB) (*The VERIS Community Database* 2021) is commonly used in cybersecurity research (Baesso Moreira et al., 2018; Y. Liu et al., 2015) and also serves as the basis for Verizon's yearly Data Breach Investigations Report (DBIR) (Bassett et al., 2021). Altogether, VCDB seems like the ideal CTI source.

As we look closer, however, problems start to emerge. VCDB is largely composed of data breach incidents collected by analysts from news reports. Although a data breach can be considered an outcome of a cybersecurity threat, it is more commonly classified as a type of threat. The European Union Agency for Cybersecurity (ENISA) is a prominent example of an institution classifying data breaches as a threat type.

ENISA publishes a yearly list of top threats (ENISA, 2020) and 'data breach' appears every year. Figure 5.1 shows a comparison of VCDB and ENISA threat rankings from 2012 to 2017. Of the 12 threats depicted, 11 appear in the ENISA top threats each year. The exception is the 'external environment threat' which was introduced by van Haastrecht, Sarhan, Shojaifar, et al. (2021). External environment threats comprise the threats resulting from third parties and suppliers interacting with an organisation. This threat category is especially relevant for SMEs, as we have seen in the proliferation of recent supply chain attacks (Browning, 2021; Lella et al., 2021). Although ENISA has not included it in their top threats, the threats making up the external environment threats do appear in their overall threat taxonomy.

To produce Figure 5.1, we analysed confirmed SME incidents included in VCDB from 2012-2017, with 2017 being the most recent year for which confirmed incidents were available. VCDB can be seen as structured open-source intelligence, but it is based on unstructured open-source intelligence. The intermediate step of structuring the original data is a time-consuming task. Thus, a common drawback of structured open-source intelligence is that it is outdated by the time it becomes available. This is problematic when the cyber threat landscape is constantly changing.

VCDB defines small businesses as having fewer than 1,000 employees, which is an exceedingly broad definition, given that it is more common to use 250 employees as the cut-off point for SMEs (European Commission, 2016). This curious SME definition is one of the reasons why using VCDB can be problematic in the SME context.

Nevertheless, we persisted in our analysis and chose to use those incidents classified as involving companies with 100 or fewer employees. Yet, as can be observed from Figure 5.1, the rankings resulting from our VCDB analysis differ from the ENISA rankings. Unsurprisingly, VCDB's focus on data breach incidents leads to a much higher ranking for the data breach threat. However,

Figure 5.1: VCDB and ENISA threat rankings compared over time. For several threats we observe large ranking differences.

many of the other threats also have ranking progressions dissimilar to ENISA's rankings.

This points to two issues with using VCDB data. Firstly, given the focus on data breach incidents, the data collected for VCDB is skewed heavily towards this threat type. This influences not only the data breach category but also all other categories, as threats that are highly correlated with data breaches will receive a higher ranking.

Secondly, since the main collection method for VCDB incidents is the scanning of news reports, the threat ranking is biased towards newsworthy threat types. Data breach incidents often appear in the news, since in many countries there is an obligation to openly report such incidents. Phishing incidents, for example, are much less likely to be reported in news articles, as companies have no incentive to communicate their occurrence.

Further issues with VCDB relate to the fact that around 82% of the SME incidents originate from the US, that the English-speaking analysts collect almost exclusively English news articles, and that the manual process of its construction results in erroneously included incidents and duplicates. Altogether, this yields a VCDB threat ranking that is unlikely to reflect the ranking obtained when having perfect knowledge of incident frequencies.

Does that mean that the VCDB is useless to SMEs? No, certainly not. By being aware of the selection bias involved in constructing the VCDB, we can still use this data as input for the prioritisation of SME cybersecurity threats. We must take care to always complement VCDB information with

other data sources, such as the ENISA rankings and expert assessments. With our approach, we hope to harness the beneficial aspects of VCDB, while taking care to avoid some of the traps associated with using its biased and outdated data.

5.3 SHARED CTI SOLUTION FOR SMES

The European Horizon 2020 project GEIGER (GEIGER Consortium, 2020) aims to develop an adaptable, dynamic, and usable application to assess and improve the cybersecurity risk level of SMEs. GEIGER achieves these goals in part by using shared threat intelligence.

Before turning to the solution we developed within the GEIGER project, let us recap what we have learned in the past two sections, to inform our solution design. We know that SMEs lack the cybersecurity knowledge and resources to perform complex tasks. Hence, they require understandable and actionable recommendations on how to improve their cybersecurity posture.

We learned that SMEs should not be seen as one homogeneous group, but rather as a heterogeneous set of enterprises with different characteristics and needs. Any cybersecurity solution for SMEs should therefore be able to adapt based on SME characteristics, to provide tailored advice.

Lastly, any cybersecurity solution needs to be updated based on changes in the cyber threat landscape. For larger enterprises, we may expect a security expert or SOC to be involved in this updating process. However, such resources are rarely available at SMEs. Therefore, our solution should incorporate an automated updating procedure facilitating adaptation to a changing threat landscape. We summarise our three requirements for an SME cybersecurity solution below:

1. The solution must provide understandable and actionable recommendations.
2. The solution should be able to adapt to different SME characteristics.
3. The solution should update automatically in response to a changing cyber threat landscape.

In the next sections, we describe how shared CTI could be the ideal prescription to meet the above requirements. The utilisation of shared CTI involves an input, a process, and an output. We cover each of these elements in the context of the GEIGER solution, starting with the input: MISP data.

5.3.1 *Input: explaining MISP*

The Malware Incident Sharing Platform (MISP) was introduced in 2016 (Wagner et al., 2016) and has risen in popularity ever since. MISP is a flexible

incident sharing platform that is compatible with STIX. The platform is supported by the Computer Incident Response Center Luxembourg (CIRCL), which explains why it is popular among many colleague Computer Emergency Response Teams (CERTs) across Europe.

MISP is a free and open-source platform for threat information sharing. MISP provides software for the sharing, storage, and correlation of IoCs related to cybersecurity incidents.

The MISP data model is composed of *events*, which usually represent threats or incidents. Events, in turn, are composed of a list of *attributes*. Examples of attributes are IP addresses and domain names. Other data types exist in MISP, such as *objects*, which allow advanced combinations of attributes, and *galaxies*, which enable deeper analysis and categorisation of events.

MISP's data model is based on a JSON schema for event exchange, allowing for the classification of objects using different taxonomies. MISP comes with predefined taxonomies and users can define taxonomies according to their needs. This allows CERTs to classify events according to their requirements, while still following accepted standards in the cybersecurity field. In Figure 5.2, we can see some examples of available taxonomies being used to classify incidents.

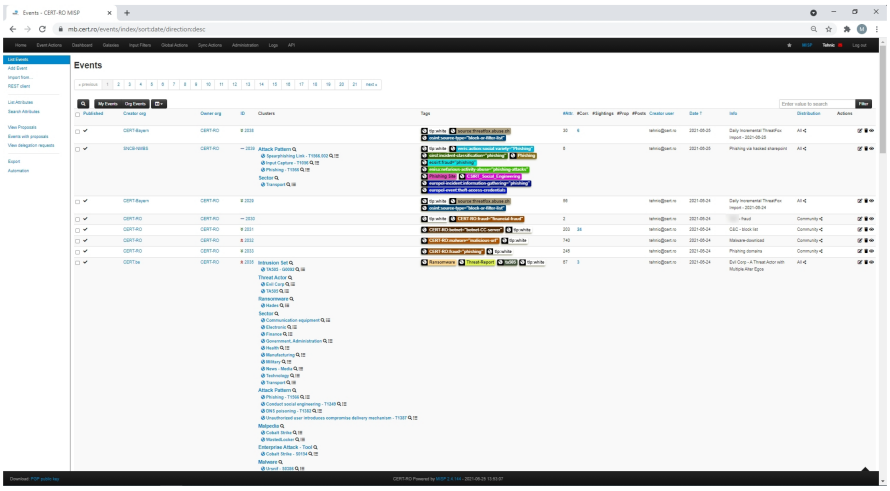


Figure 5.2: Examples of TLP:WHITE events that can be shared from CERT-RO's MISP instance to the GEIGER cloud.

CERT-RO, the Romanian CERT that is a partner in the GEIGER project, uses MISP for the collection of cybersecurity alerts from different stakeholders. To comply with its legal obligations, CERT-RO has developed a taxonomy for reporting specific events to Romanian cyberspace. All events from their sources and sensors use the CERT-RO taxonomy.

The CERT-RO MISP implementation is based on the MISP module implemented in the National Cyber Security Platform (NCSP). This platform was developed to increase CERT-RO's technical capabilities related to cybersecurity incident management and information sharing. The platform is used for the collection, processing, and dissemination of data related to cybersecurity incidents, vulnerabilities, threats, events, and artefacts, including incident notifications received by CERT-RO. Information such as malicious URLs, IPs, and file signatures are usually distributed through this module.

CERT-RO's MISP data tagged with 'TLP:WHITE' is made available to GEIGER in a feed that can be imported in the GEIGER backend component in the cloud. TLP stands for Traffic Light Protocol; a protocol created to promote the sharing of information. TLP is a set of designations used to ensure that sensitive information is shared with the appropriate audience. It employs four colours to indicate expected sharing boundaries to be applied by the recipient(s). The four colours are red (named recipients only), amber (limited distribution), green (community-wide distribution), and white (unlimited distribution). GEIGER only receives TLP:WHITE data for now. Figure 5.2 shows some examples of events shared from CERT-RO to GEIGER.

GEIGER can then use the CERT-RO CTI feed to update its solution. The technical solution used to process incoming MISP data is summarised in Figure 5.3. Information is exchanged between the GEIGER cloud storage and MISP using an information-sharing channel API. MISP JSON is shared via the information sharing channel API and temporarily stored in a raw data storage. The MISP data is then filtered to extract the information used within the GEIGER solution. The filtered information is stored in a database for processed data. Finally, the GEIGER cloud storage obtains the processed MISP events via a call to the API.

One can see that GEIGER additionally returns enriched events to MISP. Although this is a unique and useful feature in the GEIGER solution, we will not discuss it further as it falls outside of the scope of this chapter.

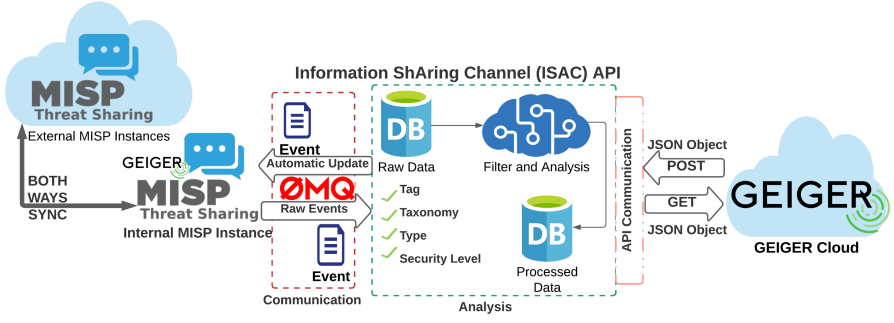


Figure 5.3: Incoming MISP data processed by the GEIGER information sharing channel and stored in the GEIGER cloud storage.

5.3.2 *Process: extracting insights from MISP data*

In our literature review, we found that researchers are starting to apply supervised machine learning (Mohasseb et al., 2020), natural language processing (NLP) (Koloveas et al., 2021; J. Zhao et al., 2020), and deep learning (Ansari et al., 2020) techniques to process shared CTI. However, we also found that applying an expert evaluation to the raw data, or using production rules, was far more popular. Of our 47 inclusions, 30 proposed the use of either an expert evaluation or production rules.

This points to the fact that shared CTI often lacks the necessary contextual information for automated reasoning, meaning some form of external knowledge has to be used during processing. This can be in a fully manual process whereby CTI is displayed and it is left to a security expert to decide what to do with the presented data. The other option is to use some form of production rules formulated by security experts a priori, whereby shared CTI can be processed automatically in production.

Of the 11 solutions in our literature review that were relevant to start-ups, digitally dependent SMEs, and digitally based SMEs, 7 used production rules in their process of turning shared CTI into usable output. This insight led us to conclude that using production rules within the GEIGER solution provides the ideal circumstances to combine expert insights with an automated, usable process for SMEs.

The GEIGER process for utilising shared CTI from MISP is depicted in Figure 5.4. We will focus on the threat prioritisation part of Figure 5.4 here, and discuss recommendation selection and the user interface in Section 5.3.3.

The threat prioritisation process proceeds as follows. First, security experts form a threat classification that is suitable for the SME target group, based on cybersecurity threat reports. In the case of GEIGER, the target audience is primarily the smallest and least digitally mature SMEs. Given their large dependence on external suppliers for IT solutions, we introduced an external environment threat representing threats from third parties and the supply chain in our classification. All other threat categories, which can be seen in Figure 5.1, appear regularly in ENISA's top threat lists. For more details on our classification, see van Haastrecht, Sarhan, Shojaifar, et al. (2021).

Next, the selected threats must be prioritised. We could choose to base prioritisation solely on the shared CTI from MISP. Yet, although MISP's threat intelligence provides a plentiful and continuous stream of data, it does not contain the information that allows us to create distinct prioritisations for different SME categories. As we outlined in our solution requirements at the start of Section 5.3, SME cybersecurity solutions must recognise the heterogeneous nature of the SME landscape. The GEIGER solution achieves this by creating different threat prioritisations for digitally dependent SMEs, digitally based SMEs, and digital enablers. Start-ups are not treated separately,

since prioritisation of threats is largely dependent on an enterprise's nature in the digital environment, rather than how long it has been in existence.

Our initial threat prioritisation was constructed based on expert insights and information from SME cybersecurity reports. Additionally, we used the insights from our VCDB analysis. We mentioned the potential issues with using VCDB data in an SME cybersecurity solution in Section 5.2. However, the analysed data can provide insights into how threat frequencies progressed over time and which threats are especially relevant to particular SME categories. An example of such an observation is that denial of service is less relevant to digitally dependent SMEs than to digitally based SMEs.

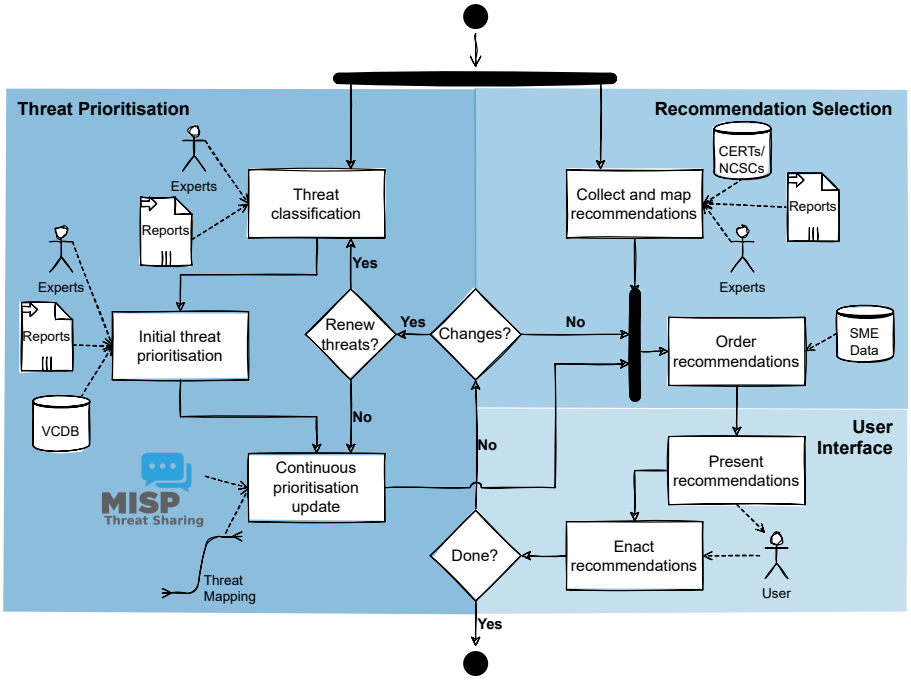


Figure 5.4: Process for turning shared CTI into actionable recommendations for SME users.

We can then use the MISP threat intelligence to continuously update our tailored threat prioritisation. However, CERT-RO's MISP taxonomies do not directly map to our ENISA-derived threat classification. Hence, we first need to use a threat mapping to map the incoming threats to the GEIGER threat classification. This mapping step is also depicted in Figure 5.3, as 'Filter and Analysis.' We can then apply our production rules to update our threat prioritisation based on the new information we receive.

To update our weights, we use an exponential smoothing approach inspired by the more advanced intermittent demand forecasting approaches known

from operations research (Nikolopoulos, 2021). In exponential smoothing, new data does not fully determine how we update our forecasts. Instead, we define a smoothing factor $\alpha \in [0, 1]$, which determines how much weight the new data receives compared to the data we already have. A lower value of α results in less weight for new data, and therefore a smoother progression over time.

The final inputs we must determine are the time intervals that we consider for updating. These intervals determine how often our algorithm should be executed, and thus how often we update threat weights in the GEIGER solution. We elected to update our weights every month, to ensure that we can respond quickly to a changing threat landscape. One might then ask: Why not update every week or every day?

We have two main reasons for not updating more frequently. Firstly, by updating very frequently we increase the influence incident outliers have on our weights. If on a particular day a large number of malware incidents are shared via MISP, this would lead to an increase in our malware weights, even though this may be unwarranted when looking at a longer period. The second reason is more practical. Users of the GEIGER application will receive recommendations based on our threat prioritisations. If we change our weights daily, users will have to deal with different prioritisations daily. From a user experience perspective, this would not be ideal.

Hence, we selected to update monthly, making the previous month the period where we consider reported incidents to be new. We label this period as t_{new} and the corresponding array of incident frequencies per threat $\mathbf{n}_{new}$. Similarly, we introduce t_{old} and $\mathbf{n}_{old}$. For these variables, we choose to look back one year, meaning incidents reported between one month ago and one year ago fall in the 'old' category.

Through the application of our exponential smoothing algorithm, we update our threat weights monthly. By updating our threat prioritisation, we ensure that the information we provide to SMEs accurately represents the current threat landscape. This allows GEIGER users to receive information on what actions they should take to counter the most pressing threats.

The process of threat prioritisation is continual, as the cyber threat landscape is ever-changing. Besides the periodic updates provided by the MISP data, we also periodically assess whether our threat classification and initial threat prioritisation should be updated.

As witnessed by the consistency in the ENISA top threats, completely new types of cybersecurity threats do not appear often. Nevertheless, given the dynamic nature of the cyber threat landscape and the constant struggle between cyber attackers and defenders, any cybersecurity solution must have controls in place to deal with major, unexpected shifts. If we observe major changes to the cyber threat landscape in our GEIGER periodic evaluations, we will repeat the complete threat prioritisation process to ensure our prioritisations are as accurate as possible.

5.3.3 *Output: providing actionable recommendations*

We observed in Section 5.2 that shared CTI solutions applying an extensive filtering process to arrive at actionable insights, are most suited to the least digitally mature SMEs. Simply providing SMEs with tailored threat prioritisations is not enough if we want to motivate them to take action. Given their lack of internally available cybersecurity expertise and resources, they need to be given clear and actionable instructions, rather than generic advice. The recommendation selection and user interface components of Figure 5.4, serve the purpose of providing SMEs with the guidance they require.

Our process starts with collecting the latest cybersecurity recommendations - sometimes termed countermeasures - from reports such as those of ENISA and the websites of national CERTs and National Cyber Security Centres (NCSCs). Many of these sources offer advice aimed specifically at SMEs.

We must then determine which recommendations apply to which threats. Luckily, many sources provide such mappings, making it relatively simple to couple recommendations to threats in the collection phase. Knowing SME characteristics such as its category, we can then order recommendations based on relevance to the SME.

Finally, we can present the ordered recommendations to the user, who can then choose to enact the recommendations they deem most relevant. Figure 5.5 shows how the GEIGER user interface presents recommendations to users.

The user receives prioritised, personalised, and actionable recommendations, without needing to first provide extensive internal data. As with any risk assessment solution, providing more data will help the SME to gain a more accurate picture of the cybersecurity risk they face. However, the user can get started without such data. This makes our approach accessible to start-ups and digitally dependent SMEs, who are in dire need of cybersecurity assistance.

5.3.4 *Practical example*

To provide insight into how the process of Figure 5.4 works in practice, we will cover a practical example in this section. The steps of our example are presented in Figure 5.6.

Recently, a malware variety termed 'Flubot' infected Android devices across Europe and Australia (Trend Micro, 2021). An increased frequency of malware incidents should be reflected in how we prioritise threats for SMEs, given that other threats are not similarly on the rise.

Figure 5.6 explains how our solution would respond to a Flubot malware wave. As the wave hits, Flubot incidents will start to appear in CERT-RO's MISP feed. The feed depicted in Figure 5.2 would change to include incident descriptions similar to the one shown in Figure 5.6.

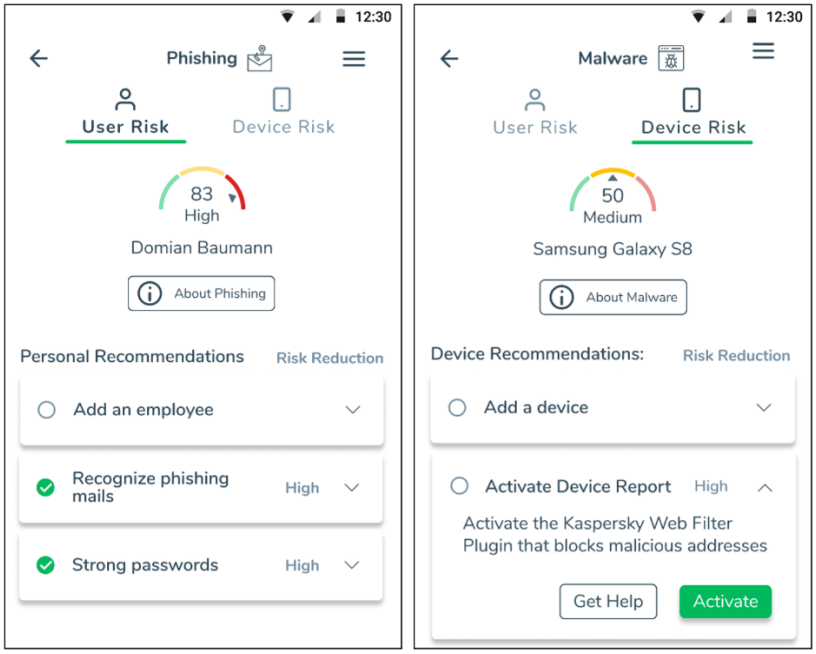


Figure 5.5: Phishing and malware recommendations shown to the user in the GEIGER user interface.

Figure 5.6: Our solution responds to the Flubot malware wave based on incoming MISP data.

The MISP data is then processed further within the GEIGER solution. Figure 5.3 showed the technical components and interactions involved in filtering MISP data and storing it in the GEIGER cloud storage. The next time the exponential smoothing algorithm is executed, the relatively high incidence of malware will cause the malware threat type to receive a higher priority. The user will be notified of a change in the prioritisation and can act accordingly. Although recommendations themselves will not be updated, the change in threat prioritisation will motivate the user to enact malware recommendations sooner rather than later.

This example highlights that just because many SMEs do not have the resources to actively monitor the cyber threat landscape, does not mean they are incapable of responding to changes. We need to construct solutions that automate the tasks SMEs are unable to perform while enabling SMEs in the tasks only they can execute. In the end, it is up to the SME to take action and implement recommendations. We, as cybersecurity experts, should do our utmost to ensure SMEs are in a position to act with confidence and determination.

5.4 DISCUSSION

At the outset of this chapter, we asked: How can shared incident information be utilised to help improve SME cybersecurity? Our literature review showed that approaches exist that could be used to help digitally based SMEs and digital enablers, but that start-ups and digitally dependent SMEs are largely left to their own devices.

We discussed how solutions building on structured external CTI show promise in helping the least digitally mature SMEs. Structured open-source intelligence also has potential, but, as our analysis of VCDB demonstrated, is likely to have biases in the data collection phase that are problematic for use in SME solutions.

Our solution using structured CTI sourced from the MISP threat sharing platform addresses the needs of the least digitally mature SMEs. In Section 5.3, we introduced three requirements for an SME cybersecurity solution, which we used to guide the design of our solution.

Our solution embeds understandable recommendations collected from CERTs and NCSCs throughout Europe in an intuitive user interface. This ensures that SMEs consider our recommendations actionable (Requirement 1). We use input from cybersecurity experts, reports, and VCDB to create a threat prioritisation tailored to an SME's category. Thus, our solution can adapt to different SME characteristics to offer tailored advice (Requirement 2). Lastly, we use incoming MISP data to continuously update our threat prioritisation, ensuring a timely response to changes in the cyber threat landscape (Requirement 3).

Our methodology and solution have their limitations. We focused our literature review on the period since the introduction of MISP in 2016. Although the last years have seen remarkable progress in the shared CTI field, it is certainly possible that we overlooked ideas for suitable solutions by restricting our timeline.

Although our application is currently complete in a prototype components implementation, its impact and relevance remain to be proven in an operational environment. We based our solution on a broad range of existing insights regarding SME cybersecurity, but it is nevertheless possible that we have overseen certain implications of using our application in the real world. An in-depth investigation of the optimal algorithm choice for updating threat weights is another future necessity.

Additionally, our solution is dependent on the continued popularity of MISP as an incident sharing platform. MISP facilitates data exchange using the STIX format, which is the de-facto standard for information exchange in the cybersecurity field. MISP, however, is not the only standard when it comes to threat sharing platforms. However, we believe in its future given the large support it receives from CERTs throughout Europe.

A final point to mention is that the validity of our solution relies on the inclusion of new cybersecurity threats in CERT-RO's MISP feed. Currently, the threats we include in our solution are all covered by one or more MISP incident types. However, if a new threat appears that is relevant to SMEs, this threat may not be represented in CERT-RO's MISP feed. This could happen if the nature of the threat makes it relevant to SMEs, yet not to CERT-RO. We believe our tight cooperation with CERT-RO and other CERTs throughout Europe offers sufficient potential for mitigation of this risk, but it is present.

5.5 CONCLUSION

Small- and medium-sized enterprises (SMEs) generally do not have the knowledge and resources to deal with cybersecurity threats. Therefore, they need to be assisted in raising their cybersecurity awareness and resilience.

A solution is to share the cyber threat intelligence (CTI) of other institutions and organisations with SMEs. After all, a problem shared is a problem halved. Yet, shared CTI is rarely used in solutions to address SME needs. Especially the least digitally mature SMEs are often overlooked.

Through reviewing the shared CTI literature, we found potential in structured, externally gathered CTI feeds to aid the most vulnerable SMEs. Our solution incorporates such an external CTI feed to continuously update threat prioritisations for SMEs. By mapping publicly available countermeasure suggestions to our prioritised threats, we can provide SMEs with actionable recommendations that are ordered by relevance.

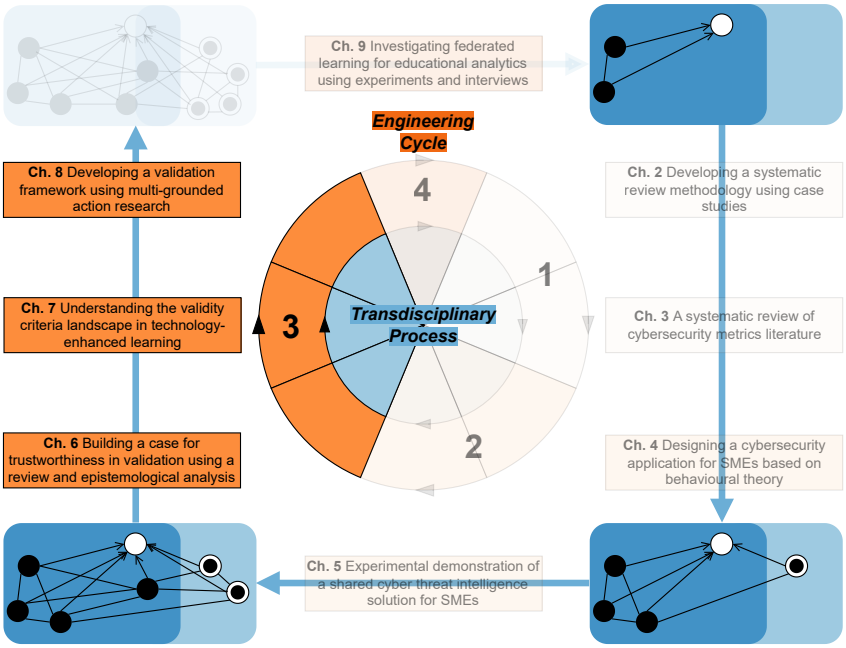
We tailored our threat prioritisations to SME characteristics, to recognise the heterogeneous SME landscape. Different SME categories deserve different

treatment, for example due to varying amounts of internal cybersecurity data being available. Our solution does not place a heavy burden on SMEs to provide internal data, making it ideally suited to less digitally mature SMEs.

Our solution is only the tip of the iceberg for what is possible with shared CTI for SMEs. In future, we will continue to develop our solution and seek to employ it in operational environments. More importantly, we hope that other researchers realise the potential of using shared CTI to help vulnerable organisations. A problem shared is a problem halved. We are well aware of the problem; it is time to start sharing.

Part III

TREATMENT VALIDATION



6

EMBRACING TRUSTWORTHINESS AND AUTHENTICITY IN VALIDATION

Learning analytics sits in the middle space between learning theory and data analytics. The inherent diversity of learning analytics manifests itself in an epistemology that strikes a balance between positivism and interpretivism, and knowledge that is sourced from theory and practice. In this chapter, we argue that validation approaches for learning analytics systems should be cognisant of these diverse foundations. Through a systematic review of learning analytics validation research, we find that there is currently an over-reliance on positivistic validity criteria. Researchers tend to ignore interpretivistic criteria such as trustworthiness and authenticity. In the 38 papers we analysed, researchers covered positivistic validity criteria 221 times, whereas interpretivistic criteria were mentioned 37 times. We motivate that learning analytics can only move forward with holistic validation strategies that incorporate "thick descriptions" of educational experiences. We conclude by outlining a planned validation study using argument-based validation, which we believe will yield meaningful insights by considering a diverse spectrum of validity criteria.

The contents of this chapter are based on: van Haastrecht, M. Brinkhuis, Peichl, et al. (2023). Embracing Trustworthiness and Authenticity in the Validation of Learning Analytics Systems. In Proceedings of the 13th International Learning Analytics and Knowledge Conference.

6.1 INTRODUCTION

In a recent survey among learning analytics experts (R. Ferguson et al., 2019), validity was ranked as the third-most important theme relating to the future of learning analytics, behind power (i.e., control over data) and pedagogy. R. Ferguson et al. (2019) state that validation approaches should always take "context into account when reporting results". Recognising that each instructional context is different is seen by Gašević, Dawson, et al. (2016) as a prerequisite for an acceptable validation strategy. Kitto et al. (2018) agree, arguing that validation must address both positivistic (e.g., performance metrics) and interpretivistic (e.g., student experience) elements. They conclude that "work on developing new validation criteria that emphasise learning outcomes" is vital. This conclusion is in agreement with the experts in R. Ferguson et al. (2019), who state that "research in this space should be tied to pedagogical outcomes."

Thus, validation is a critical topic for learning analytics research. There is agreement that validation should go beyond performance metrics and that an additional emphasis on learning outcomes would help to yield a contextualised approach. Yet, there is little consensus on which validity criteria are essential in learning analytics research. In a recent special issue on the potential links between learning analytics and educational assessment, Gašević, Greiff, et al. (2022) raised the concern that "existing learning analytic methods do not meet all of the criteria" for validation we encounter in educational assessment. However, Gašević, Greiff, et al. (2022) do not discuss to which criteria they are referring. The learning analytics literature lacks an in-depth analysis of the validity criteria that are currently in use and the criteria that deserve emphasis. We will address this gap in this chapter.

With the previous paragraphs in mind, we formulate the following main research question and sub-questions:

- **RQ**: Which validity criteria should be considered in a contextualised validation strategy for learning analytics systems?
  - **RQa**: Which validity criteria have emerged in the learning analytics domain that emphasise learning outcomes?
  - **RQb**: How has learning analytics validation research incorporated interpretivistic perspectives that recognise contextual differences?

Through an analysis of the epistemological foundations of learning analytics (Section 6.2) and a systematic review of the learning analytics validation literature (Section 6.3), we will construct an overview of emerging validity criteria to answer **RQa**. An in-depth analysis of our systematic review results (Section 6.4) will help us in answering **RQb**. We discuss the implications for our main research question in Section 6.5 and conclude in Section 6.6.

6.2 BACKGROUND: EPISTEMOLOGY AS A FOUNDATION FOR VALIDATION

How we approach validation depends on our underlying epistemology, specifically relating to our view on the concept of truth. A purely interpretivist researcher will attach little value to performance metrics when validating since they reject the concept of objective truth in social contexts. Similarly, positivist researchers are unlikely to engage in what Geertz (1973) termed "thick description" of social contexts, as they believe in the generalisability of more efficiently obtainable quantitative evidence. We posit that learning analytics epistemology is positioned in the middle space between interpretivism and positivism. In this section, we will provide further intuition for this observation and motivate that the axis of truth is not the only epistemological axis relevant to building a solid foundation for validation.

Pragmatism is one of the cornerstones of today's learning analytics literature. As envisioned by Dewey (1931), pragmatism takes a moderate position in the interpretivism versus positivism debate. Kuhn (1962) describes the scientific process as "a process whose successive stages are characterised by an increasingly detailed and refined understanding of nature." A process of moving "from primitive beginnings," yet not "towards anything." This contradicts the positivist view that the scientific method enables us to consistently hone in on truths and thereby expand our knowledge. Dewey (1938) avoids the term knowledge altogether, preferring "warranted assertability." This phrase connects the past (warranted) and the future (assertability). Dewey's pragmatism, therefore, blends views that aim to build from a common past (interpretivism) with those that aim to move towards a common future (positivism).

However, the axis of truth is not the only relevant epistemological axis when laying the foundations for validation. Pragmatists claim that "our conception of some given thing is bound up in our understanding of its practical application" (Knight et al., 2014). Not only a definition of what constitutes knowledge is crucial, but also a consideration of possible sources of knowledge. Pragmatism posits that practical use should be the primary source of knowledge, which juxtaposes it with rationalism which states that theoretical reasoning is the summum bonum when it comes to knowledge gathering. Wise et al. (2016) propose a similar classification regarding learning analytics design knowledge. They state that design knowledge can originate from the design process, which is guided by theory, and from the implementation process, which is coupled with the introduction of learning analytics in the learning environment.

Dewey helped develop a version of pragmatism, known as transactionalism, that emphasises contextual interactions as a vital source of knowledge (Dewey and Bentley, 1949). Transactionalism merges ideas from pragmatism and constructivism, with Dewey's version of pragmatism being considered "as the most important precursor for social constructivism" (Reich, 2007). Social constructivism is the variant of constructivism most often encountered

in learning analytics research today. Social constructivists argue "that learners arrive at what they know mainly through participating in the social practices of a learning environment" (Woo and Reeves, 2007). Social constructivism focuses on meaningful interactions in authentic contexts. However, in today's world, many educational interactions involve technological assistance. Although there is a role for social constructivism in technology-enhanced learning (Woo and Reeves, 2007), its focus on social interactions as the primary source of knowledge makes it ill-suited to assess the consequences of today's socio-technical systems. Siemens (2004) aimed to solve this issue with connectivism.

Connectivism is perhaps the philosophical stance most closely associated with learning analytics. Connectivism is similar to social constructivism, but it reserves an explicit place for "learning that occurs outside of people (i.e. learning that is stored and manipulated by technology)" (Siemens, 2004). Connectivism, like learning analytics itself, states that theory is a valid source of knowledge. This brings us, finally, to the place that learning analytics epistemology occupies within the epistemological plane of Figure 6.1. We propose that learning analytics epistemology is positioned in the middle space between positivism and interpretivism on the axis of truth, but also in the middle space between theory and practice on the knowledge source axis.

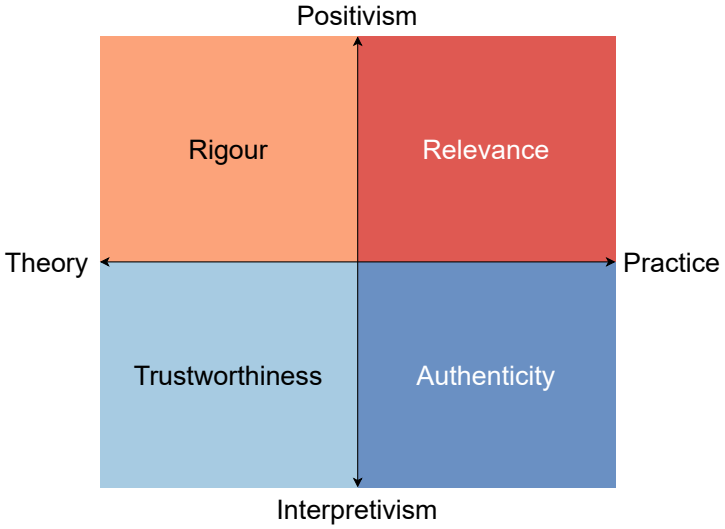


Figure 6.1: Our epistemological plane of validity, divided into four quadrants. Learning analytics occupies the middle space between positivism and interpretivism (the truth axis), and the middle space between theory and practice (the knowledge source axis).

Figure 6.1 introduces the overarching terms we use within this chapter to refer to the four quadrants created by the axes of our epistemological plane.

In the positivistic tradition, it is common to distinguish between rigour and relevance in research (Hevner et al., 2004). Rigour is connected to theory as a source of knowledge and can be "achieved by appropriately applying existing foundations and methodologies" (Hevner et al., 2004). Research is relevant when it addresses "the problems faced and the opportunities afforded by the interaction of people, organisations, and information technology" (Hevner et al., 2004). On the side of interpretivism, Guba (1981) proposed the concept of trustworthiness as a parallel to rigour. Lincoln and Guba (1986) later introduced authenticity as a more practice-oriented validity conceptualisation, noting that "conventional criteria refer only to methodology and ignore the influence of context."

Figure 6.1 only presents four overarching validity quadrants, providing an incomplete answer to **RQa** on emerging validity criteria. Many more criteria are considered in the learning analytics literature, each with its own place within the epistemological plane. To investigate which criteria are considered and whether specific areas of the epistemological plane are underrepresented, we conducted a systematic review of the learning analytics validation literature.

6.3 METHODOLOGY

For our systematic review, we queried three databases: ACM Digital Library, Web of Science, and PubMed. We searched for all papers with abstracts containing the phrase 'learning analytics' and either 'validation' or 'validity'. After deduplicating the query results, 83 papers remained. Of these papers, 21 formed the initial set of inclusions after excluding work that did not discuss validation or was unrelated to the field of learning analytics (as defined by SoLAR (Society for Learning Analytics Research (SoLAR), 2022)). For each of these 21 included papers, we scanned all the references and citations to find potential new inclusions. This process is known as 'snowballing' and is a recommended step in systematic review methodologies (van Haastrecht, Sarhan, Yigit Ozkan, et al., 2021). The snowballing phase resulted in a further 17 inclusions, meaning our final set comprised 38 papers.

Before proceeding to analyse our inclusions, we identified four papers which would allow us to construct a holistic set of potential validity criteria. We first looked towards educational measurement (sometimes referred to as educational assessment). Educational measurement is a field where validity considerations naturally take centre stage, and several learning analytics researchers have argued that we should strengthen the bond with this field (Gašević, Greiff, et al., 2022). The argument-based validation approach of Kane (2013b) has been influential in the educational measurement and learning analytics fields in recent years (Douglas et al., 2020; Milligan, 2018). Kane (2013b) stresses the importance of addressing traditional validity criteria such as rigour, construct validity, content validity, and criterion validity. However,

Kane's framework also recognises that theoretical considerations alone are insufficient, and that validation must investigate how results are used in practice. Kane captures this idea in the concept of consequential validity.

The fields of design science and information systems offer a second source of inspiration in the validity considerations made by learning analytics researchers. Mingers and Standing (2020) provide an extensive overview of the validation literature in these fields, while highlighting the importance of the interpretivistic perspective. The criteria external validity (sometimes termed generalisability), internal validity, reliability, replicability, and statistical validity occupy the rigour quadrant. Mingers and Standing (2020) additionally propose consistency (relevance quadrant) and elegance (authenticity quadrant) criteria.

Our third external source of validity terminology is the seminal interpretivistic work of Lincoln and Guba (1986). Their paper introduced the concept of authenticity as a counterbalance to trustworthiness. Lincoln and Guba (1986) discuss various dimensions of trustworthiness that parallel positivistic criteria: confirmability (related to replicability and content validity), credibility (internal validity), dependability (reliability), and transferability (external validity). They additionally discuss several dimensions of authenticity, but we select to include authenticity as a single criterion in this chapter as this is generally how the construct is viewed in learning analytics research. Lastly, Lincoln and Guba (1986) introduce fairness as a vital consideration during validation.

Finally, certain validity considerations are quite unique to the learning analytics field. To provide sufficient coverage of these validity criteria, we looked towards the work of Ali et al. (2012). They propose a diverse selection of validity criteria covering the relevance quadrant (relevance, actionability, understandability, usability, and usefulness) and the authenticity quadrant (meaningfulness and parsimony/simplicity).

6.4 RESULTS

Figure 6.2 depicts the assembled validity criteria within their respective quadrants. Criteria are positioned according to how they are defined and treated in the literature, thereby acting as a Learning Analytics Validation Assistant (LAVA). Researchers can use LAVA to determine whether the validity criteria they are considering are sufficient and appropriate for their epistemological stance. A criterion's quadrant is determined by how it is defined in one of the four core papers mentioned in the previous section. The exact placement of a criterion within a quadrant should not be interpreted as an indisputable truth. Rather, we positioned criteria relative to each other based on how they were treated and measured in the learning analytics literature.

Figure 6.2 additionally visualises the prevalence of the validity criteria in our included papers. In 38 inclusions, a total of 258 validity criteria were

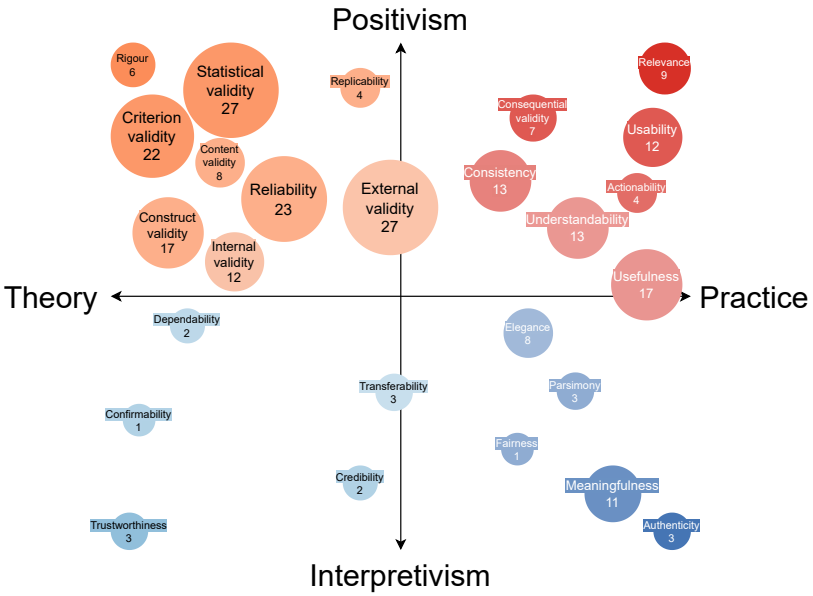


Figure 6.2: The Learning Analytics Validation Assistant (LAVA), depicting the prevalence of validity criteria observed in the learning analytics literature. Terms are positioned along the two axes (truth: positivism versus interpretivism; knowledge source: theory versus practice) of our epistemological plane.

discussed. Criteria in the rigour quadrant were mentioned 146 times (56.6%), in the relevance quadrant 75 times (29.1%), in the trustworthiness quadrant 11 times (4.3%), and in the authenticity quadrant 26 times (10.1%). Hence, researchers covered positivistic criteria 221 times, whereas interpretivistic criteria were mentioned only 37 times.

Statistical validity and external validity are the criteria mentioned most often within our inclusions. For statistical validity, we noticed that most papers focus on statistical significance, whereas Saqr and López-Pernas (2021) point out that researchers should additionally consider effect size. External validity is another problematic criterion within learning analytics research. Of the 27 times external validity was mentioned in one of our inclusions, 24 times the authors concluded that the external validity of their study was lacking. We observed a similar pattern with the interpretivistic counterpart to external validity: transferability. Of the three times transferability was considered, the authors stated on two occasions that more work was necessary to assess the transferability of their results.

Figure 6.2 provides an answer to **RQa**: Which validity criteria have emerged in the learning analytics domain that emphasise learning outcomes? Criteria on the 'practice' half of the diagram relate to outcomes of the learning process. Criteria positioned on the extreme right of the theory-practice axis correspond to an advanced internalisation of learning analytics outcomes. Learning analytics researchers evidently attach importance to relevant, usable, actionable, and useful solutions. Additionally, several papers recognised that authentic, meaningful learning experiences are not simply a luxury, but a goal to strive for.

Table 6.1 lists the combinations of validity criteria quadrants observed in our inclusions. Four out of 38 papers covered criteria from all four quadrants. Papers tended to consider criteria from at least two quadrants, with only one inclusion not covering a criterion from the rigour quadrant. Conversely, the criteria in the trustworthiness quadrant, along with parsimony, authenticity, and fairness, are mentioned least often. Many of these criteria can only be assessed through "thick descriptions" of social contexts (Geertz, 1973), possibly pointing to barriers to engaging in such activities within learning analytics research. Moreover, although meaningfulness was discussed in 11 papers, only one of these papers conducted qualitative interviews during validation. All other papers used either quantitative data analysis or structured questionnaires in their evaluation. Concerning **RQb**, we can conclude that although interpretivistic validity criteria are considered in learning analytics research, their treatment is often too superficial to provide in-depth insight into contextual learning experiences.

Table 6.1: Combinations of the four validity criteria quadrants (rigour, relevance, trustworthiness, and authenticity) observed in the 38 inclusions of our systematic review, sorted by number of related inclusions. Only observed combinations are listed.

QUADRANT COMBINATION	RELATED INCLUSIONS
Rigour, relevance	Berman and Artino (2018), Chaparro-Peláez et al. (2020), Dourado et al. (2021), Effenberger and Pelánek (2021), Fincham et al. (2019), Galaige et al. (2018), Giannakos et al. (2015), Howell et al. (2018), Saqr, Viberg, et al. (2020), and Tabuenca et al. (2015)
Rigour, relevance, authenticity	Alonso-Fernández et al. (2019), Ifenthaler and Widanapathirana (2014), Kärner et al. (2021), Muñoz et al. (2020), Pardo et al. (2015), Park and Jo (2019), Saqr and López-Pernas (2021), Whitelock-Wainwright et al. (2020), Ye and Pennisi (2022), and Zheng et al. (2022)
Rigour	Bitner et al. (2020), Jo et al. (2014), Maldonado-Mahauad et al. (2018), Matcha et al. (2020), and Prat and Code (2021)
Rigour, authenticity	Chejara et al. (2021), Fan, Lim, et al. (2022), D. Kim et al. (2016), Kizilcec et al. (2017), and Sinha et al. (2014)
Rigour, relevance, trustworthiness, authenticity	Ali et al. (2012), Valle et al. (2021), Wise et al. (2016), and Yoo and Jin (2020)
Rigour, relevance, trustworthiness	Cerro Martínez et al. (2020), Saqr, Fors, et al. (2018), and Winne (2020)
Relevance	Gañán et al. (2017)

6.5 DISCUSSION

Our results lead to three main findings related to the learning analytics validation literature, which we will cover in this section.

6.5.1 *Troubling external validity*

Learning analytics researchers seem to have a troubling relationship with external validity. Together with statistical validity, external validity was the criterion mentioned most often in our inclusions. Yet, 24 out of the 27 papers that mention external validity conclude that there are limitations to the generalisability of their results. At times, the limited scale of studies is listed as the cause for generalisability concerns (e.g., (Chaparro-Peláez et al., 2020; Tabuenca et al., 2015; Wise et al., 2016; Yoo and Jin, 2020)). Elsewhere, researchers provide a general warning that more research is necessary should one want to generalise the results (e.g., (Effenberger and Pelánek, 2021; D. Kim et al., 2016; Saqr, Fors, et al., 2018)). Transferability, the interpretivistic parallel of external validity, suffers from the same issue. Researchers state that results could be transferred to other contexts, but that more research is required to confirm this claim (Ali et al., 2012; Cerro Martínez et al., 2020).

The reader should not interpret the previous paragraph as a critique of the cited research. If there are limitations to the generalisability of findings, these should be mentioned. However, we should avoid a situation in the learning analytics field where generalisability becomes an afterthought that can always be left for future work. External validity and transferability are valued validity criteria that should guide learning analytics research a priori, not a posteriori.

Replication studies that aim to understand the validity of learning analytics solutions in new contexts should receive more attention.

6.5.2 *A need for thick descriptions*

We noted in Section 6.4 that even papers that recognise interpretive validity criteria (e.g., meaningfulness) often resort to quantitative methods during validation. Geertz (1973) believes that the analysis of social culture and context requires qualitative methods "in search of meaning" rather than quantitative methods "in search of law." In other words, we require "thick descriptions" of the educational contexts being considered in learning analytics research. Thick descriptions that cannot be obtained through data analysis or questionnaires, but that require qualitative methods.

The advantages of using qualitative methods go beyond a deeper understanding of the educational context. As Guba (1981) recognises, "to determine the extent to which transferability is probable, one needs to know a great deal about both the transferring and receiving contexts." Guba (1981) states that thick descriptions are essential if we wish to achieve transferable results. Thus, thick descriptions provide deeper insight into interpretivistic validity criteria and concurrently act as a catalyst in facilitating generalisable learning analytics research. Researchers looking to produce more generalisable results will benefit from employing qualitative research methods such as qualitative interviews and action research.

6.5.3 *The potential of argument-based validation*

To conclude this section, we will discuss a validation approach uniquely suited to facilitate the diverse validity criteria and research methods covered in this chapter: argument-based validation. Kane (2013b) originally introduced this approach in the educational measurement field. Gašević, Greiff, et al. (2022) argue that learning analytics research can profit from the vast validity experience within educational measurement and psychological assessment, and argument-based validation has started to see use within the learning analytics domain (Douglas et al., 2020; Milligan, 2018).

In general, research uses inferences to make warranted claims based on data. Argument-based validation proceeds by constructing arguments to provide evidence for the assertability of these claims. Once evidence has been assembled in structured arguments, we assess the validity of the overall inference chain. The benefit of this approach is that it gives a balance of flexibility and structure, allowing researchers to recognise "legitimately diverse arguments" (Addey et al., 2020) while avoiding the open-ended nature of validation. The original framework of Kane (2013b) has been extended to allow for an increased focus on practical consequences (Hopster-den Otter et al., 2019) and to

explicitly address fairness in artificial intelligence (AI) enhanced assessments (Huggins-Manley et al., 2022). Argument-based validation is a promising avenue for learning analytics researchers looking to address diverse validity criteria and produce rigorous, relevant, trustworthy, and authentic results.

6.6 CONCLUSION AND FUTURE WORK

Within this chapter, we have investigated which validity criteria should be considered in a contextualised validation strategy for learning analytics systems. We proceeded by first analysing the epistemological foundations of learning analytics research, concluding that learning analytics epistemology is positioned in the middle space between positivism and interpretivism and between theory and practice. We then conducted a systematic review to uncover which types of validity criteria are employed by learning analytics researchers. We visualised the results to create a Learning Analytics Validation Assistant (LAVA).

We uncovered an over-reliance on positivistic criteria. Interpretivistic criteria that were covered (e.g., meaningfulness), were often investigated using quantitative rather than qualitative methods. In Section 6.5, we analysed the LAVA results and delineated a need for more focus on "thick descriptions" of educational experiences. Such thick descriptions help to foster a deeper understanding of the context being studied and can act as a catalyst in facilitating generalisable research.

In future work, we will apply our LAVA insights within an educational research project. As suggested in Section 6.5.1, we intend to employ an argument-based validation approach incorporating diverse arguments and validity criteria. We recognise that we are bound to encounter limitations in our future work and want to stress that no single approach can function as a validation panacea. Nevertheless, we believe that LAVA can stimulate researchers to evaluate whether their validity criteria are sufficient and appropriate for their epistemological stance.

7

VALIDITY CRITERIA FOR TECHNOLOGY-ENHANCED LEARNING

Technological aids are ubiquitous in today's educational environments. Whereas much of the dust has settled in the debate on how to validate traditional educational solutions, in the area of technology-enhanced learning (TEL) many questions still remain. Technologies often abstract away student behaviour by condensing actions into numbers, meaning teachers have to assess student data rather than observing students directly. With the rapid adoption of artificial intelligence in education, it is timely to obtain a clear image of the landscape of validity criteria relevant to TEL. In this paper, we conduct a systematic review of research on TEL interventions, where we combine active learning for title and abstract screening with a backward snowballing phase. We extract information on the validity criteria used to evaluate TEL solutions, along with the methods employed to measure these criteria. By combining data on the research methods (qualitative versus quantitative) and knowledge source (theory versus practice) used to inform validity criteria, we ground our results epistemologically. We find that validity criteria tend to be assessed more positively when quantitative methods are used and that validation framework usage is both rare and fragmented. Yet, we also find that the prevalence of different validity criteria and the research methods used to assess them are relatively stable over time, implying that a strong foundation exists to design holistic validation frameworks with the potential to become commonplace in TEL research.

The contents of this chapter are based on: van Haastrecht, Haas, et al. (2024). Understanding Validity Criteria in Technology-Enhanced Learning: A Systematic Literature Review. Computers & Education.

7.1 INTRODUCTION

Validation in its most general sense involves evaluating evidence regarding specific claims, to assess the plausibility of these claims (Kane, 1992). Validity is a multi-faceted concept, with different validity criteria being more or less relevant in different contexts. When Cronbach and Meehl (1955) and Messick (1989) were emphasising the need for well-defined validity criteria in the second half of the twentieth century, the immense influence technology would come to have on our daily lives had yet to materialise. Concepts such as construct validity, criterion validity, and content validity seemed to cover the most vital aspects of validity in educational measurement and assessment (Kane, 1992). Then came the introduction of the varied assortment of technologies that exist today to enhance our educational environments. Students can now collaboratively improve their problem solving skills using online platforms (Stadler et al., 2020) and we can use learning analytics to understand the behaviour of students on the other side of the world in Massive Open Online Courses (MOOCs) (Douglas et al., 2020). This raised the question: can we continue to use traditional validity criteria in these decidedly non-traditional contexts?

The short answer to this question is no: we cannot simply apply old validity criteria to a new age. We need to recognise that validity argumentation must adapt when we switch from traditional classroom settings to complex, interactive environments at scale (Mislevy, 2016). External validity, commonly referred to as generalisability, is a typical criterion that we should be mindful of in technology-enhanced learning (TEL) settings. Technologies tend to abstract away the context of the learner and present educators with student data that can at best provide a summary of the actual learner context. Generalisation arguments rely on some form of comparability between one context and the next, and when students use their own devices and software, the comparability claim is challenged (Wools, Molenaar, et al., 2019). With respect to a validity criterion such as authenticity, we can use virtual reality and simulations to create more authentic educational experiences (Wools, Molenaar, et al., 2019), but in cases where technology abstracts away student behaviour, it can become difficult to assess how authentic these educational experiences really are (van Haastrecht, M. Brinkhuis, Peichl, et al., 2023).

In a special issue examining possible links between learning analytics and assessment, the editors stated that "the field still needs a clear theoretical framework to guide the consideration of validity" (Gašević, Greiff, et al., 2022, p. 4). Recent work looking to deepen the connection between these two fields highlights that future research can improve the validity of learning construct interpretations by combining insights from different data sources (Raković et al., 2023). Likewise, several systematic reviews have stressed the need for a coherent, comprehensive framework to aid with the evaluation of TEL environments (Clunie et al., 2018; Erdt et al., 2015; Heil and Ifenthaler, 2023).

We need to clarify the current landscape of validity criteria in TEL if we are to facilitate rigorous validation in the future. Without consensus on which validity criteria could possibly be examined, we cannot expect researchers to make the right considerations.

Researchers have started to theorise what validation should look like in the technology-enhanced world of today. A common factor among novel views regarding validation is the necessity to recognise diverse perspectives (Addey et al., 2020) and diverse epistemologies (van Haastrecht, M. Brinkhuis, Peichl, et al., 2023). Recognising diverse perspectives does not exclude the possibility of reaching a common theoretical foundation. We can recognise that evidence for certain validity criteria (e.g., authenticity) may be sourced from qualitative methods applied in practice whereas evidence for other criteria (e.g., statistical validity) may result from quantitative methods based on theory, while still identifying epistemological patterns in validity criteria relations that can serve as a basis for holistic validation frameworks.

By combining insights on the validity criteria considered in TEL research, how they are defined and measured, how their prevalence has evolved over time, and how criteria relate in an epistemological sense, we can take a first step in addressing the current gaps in the literature relating to TEL validation. We aim to gain these insights in this paper by conducting a systematic review of TEL literature, to uncover how researchers have dealt with the challenging nature of TEL validation. There is, to our knowledge, no systematic review that investigates the validity arguments TEL researchers rely on to defend their conclusions. By collecting details on how validity criteria were defined and measured, we will answer the following research questions:

- **RQ**: How can we characterise the landscape of validity criteria used in TEL research?
  - **RQa**: Which validity criteria are considered in TEL research, how are they defined, and how are they measured?
  - **RQb**: How has the prevalence of different validity criteria in TEL research evolved over time?
  - **RQc**: What epistemological patterns do we observe in the connections between validity criteria in TEL research?

In what follows, we will first discuss earlier work on TEL validation and reviews of TEL literature (Section 7.2). We will then outline our systematic review methodology in Section 7.3 and present our results in Section 7.4. We discuss and interpret our results in Section 7.5, also covering some of the limitations of our methodology. Section 7.6 concludes and outlines several interesting areas for future research.

7.2 BACKGROUND

In this section, we outline the challenges posed to existing validity arguments when technology enters the picture, the solutions that have been proposed, and current open problems. Additionally, we discuss previous systematic reviews of TEL research, demonstrating how our work contributes to the existing literature.

7.2.1 *Validation of technology-enhanced learning*

Bennett and Bejar (1998) recognised over 25 years ago that the introduction of technology into our learning environments necessitated a different approach to validation. They argued that validation cannot be complete unless the underlying rationales supporting design decisions are adequately explained. Kane (1992, 2013b), whose argument-based approach to validation (Kane, 1992, 2013b) is considered to be dominant in educational assessments (Addey et al., 2020), recognised that the introduction of technology implied that different elements in the validity argument now required emphasis (Clauser et al., 2002). In the argument-based approach to validation, an inference chain is constructed pertaining to the design in question, whereby evidence is collected to inform arguments supporting the validity of each step in the inference chain. If we do not deal with novel threats to validity, such as generalisability issues caused by students using personal devices and software (Wools, Molenaar, et al., 2019), we risk weakening the links of the inference chain and undermining the trustworthiness of TEL systems (Aloisi, 2023). The challenges posed by the introduction of novel technologies have led to the conclusion that adapted validation frameworks are required to deal with our adapted world (Mislevy, 2016).

Several adapted validation frameworks have been proposed in recent years that build on the argument-based approach. Zhai et al. (2021) introduced a validity inferential network to better incorporate the impact of machine learning on today's educational assessments. Huggins-Manley et al. (2022) similarly focus on how assessments enhanced with artificial intelligence should be validated, taking a specific interest in fairness. In van Haastrecht, M. J. S. Brinkhuis, Wools, et al. (2023), a validation framework for e-assessment solutions is proposed that combines traditional insights on validity from the educational domain with information systems validity theory. However, these frameworks are rarely employed within general TEL research outside of educational measurement. This is evidenced by a recent systematic review where the authors stated that, to the best of their knowledge, no such frameworks existed (F. L. da Silva et al., 2023).

Where validation generally covers the full research cycle, including research methodology and design methods, evaluation tends to focus on the artefact produced by research and how it is used. Evaluation is more common in

TEL research than validation, but evaluation strategies are generally not comprehensive in nature. A review of TEL literature found that the majority of studies cover one or two educational aspects when evaluating the use of TEL solutions, leading the authors to question "whether educators are evaluating the use of technology in education from a holistic perspective" (Lai and Bower, 2019, p. 38). These findings were largely confirmed in a follow-up study where Lai, Bower, et al. (2022) asked educational technology experts which dimensions should be considered when evaluating TEL solutions. Although the experts could agree to a large extent that learning outcomes and technological aspects should be considered during evaluation, aspects such as design and behaviour were only considered relevant by a minority. The study concludes that theories used in TEL evaluation studies "do not comprehensively account for all dimensions of educational technology use" (Lai, Bower, et al., 2022, p. 752). The review authors describe how they validated their questionnaire, but do not discuss the relationship between the validation and evaluation of TEL research, pointing to the disconnect between current TEL studies and the body of knowledge on validity theory from educational measurement.

That is not to say that TEL researchers are unaware of approaches such as argument-based validation. In fact, several recent works in the area of learning analytics have stressed the potential of argument-based validation to yield more holistic evaluations (Gašević, Greiff, et al., 2022; van Haastrecht, M. Brinkhuis, Peichl, et al., 2023). We have seen applications of the argument-based approach to validation in studies concerning MOOC assessment (Douglas et al., 2020), asynchronous writing tasks (T. Chen, 2022), and eye-tracking solutions for the assessment of data literacy (F. Chen et al., 2023). However, these studies use the traditional argument-based approach, rather than frameworks adapted to suit technology-enhanced environments. Combined with the general lack of comprehensive evaluations, it is evident that TEL validation is still in its infancy (Rodríguez-Triana et al., 2017).

Employing adapted validation frameworks is required to move TEL validation from infancy to maturity, but using such frameworks passively is not sufficient. We must also continuously develop these frameworks to align with the novel learning process data that becomes available due to technological advancement (Fan, van der Graaf, et al., 2022; Raković et al., 2023). Goldhammer et al. (2021) argue that a complete validity argument requires thought about process indicators right from the start of the design phase. Yet, Zumbo et al. (2023) find that there are currently no holistic validation frameworks that adequately deal with TEL process data. Furthermore, only adapting to technological advancement is insufficient. We need to actively ensure our approaches to validation appreciate the human element in the face of increasing technological influence. Future validation approaches should leave room for legitimately diverse arguments (Addey et al., 2020) that consider qualitative criteria such as fairness and trustworthiness (van Haastrecht, M. Brinkhuis,

Peichl, et al., 2023). Only then can we truly claim that TEL validation has matured.

7.2.2 *Systematic reviews of technology-enhanced learning*

Before moving forward with our review, we should ask: How have earlier systematic reviews addressed the topics of validation and evaluation? Verbert et al. (2012) review recommender systems for TEL and retrieve information on whether studies evaluated learning efficiency/effectiveness, accuracy, usefulness, and usability. They find that some studies perform no evaluation at all and that the majority of studies only consider one or two criteria during evaluation. These findings lead Verbert et al. (2012) to conclude that more comprehensive evaluation studies are needed with a more structured approach. Yet, they do not detail what such an approach may look like and which further criteria might be needed to arrive at a comprehensive evaluation. Erdt et al. (2015) similarly review recommender systems for TEL and focus explicitly on evaluation. Of the 235 studies they include, 95 performed no evaluation. The authors suggest that we need to consider evaluation from the earliest design stage and that we should use evaluation frameworks to standardise the evaluation process. However, like Verbert et al. (2012), the authors focus on evaluation, not validation. Evaluation is mostly geared at answering questions about designed artefacts and how they are used, whereas validation also critically examines the research and design methods that produced an artefact. The fixation of evaluation approaches on outcome over process naturally produces more insights regarding implementation than early design stages.

Later TEL reviews maintain this focus on outcomes. Boyle et al. (2016) review the impacts and outcomes of computer games and serious games. Rodríguez-Triana et al. (2017) review blended TEL environments, finding that usefulness and usability are the most commonly incorporated constructs in evaluations, and concluding that their findings are illustrative of a relatively young field. Clunie et al. (2018) ask whether studies investigating the efficacy of TEL resources are comprehensive, in the sense that studies go beyond measuring learner impact to also consider institutional impact. The authors find that no study considered the institutional perspective, and mention the need for "robust evaluation strategies that can provide answers to the why, how, and when questions" (Clunie et al., 2018, p. 315). Lai and Bower (2019) confirm these findings, showing that just 1.4% of studies consider the institutional environment during TEL evaluation. As with the other reviews we have discussed, neither of these reviews mention the possibility that a focus on validation rather than evaluation could be the solution.

In a 2020 tertiary review of 73 systematic reviews, Lai and Bower (2020) provide further evidence of the lack of consensus in TEL evaluation. Using the eight dimensions of evaluation from their earlier work (Lai and Bower, 2019) - including learning, behaviour, design, and the institutional environment -

they find that no systematic review covered more than five dimensions. The authors assert that there is room for a systematic review taking a broader perspective of evaluation "to more comprehensively understand the effects of using technology in education" (Lai and Bower, 2020, p. 253). In other words, at the time of the tertiary review, there was a need for the type of systematic review we are performing within this work.

Since 2020, we have seen various systematic reviews in the area of TEL, but none have tackled the broader perspective that Lai and Bower (2020) call for. Some of these reviews focus on a specific criterion such as generalisability (Abdulrahaman et al., 2020) or usability (Law and Heintz, 2021), thereby not offering a comprehensive overview. Others consider multiple criteria, but do not employ a specific framework or set of evaluation dimensions (Bond et al., 2020; F. L. da Silva et al., 2023; Heil and Ifenthaler, 2023). Further reviews take a different perspective entirely, and consider what drives the adoption of learning technologies (Q. Liu et al., 2020), evaluate how effective workshops are that prepare teachers for TEL (Ahadi et al., 2021), or analyse the survey instruments used to evaluate the integration of new technologies (Consoli et al., 2023).

We can conclude that the gap in the literature identified by Lai and Bower (2020) has not yet been addressed. We still require a systematic review that provides a comprehensive overview of the landscape of evaluation and validity criteria in TEL research. Furthermore, we have seen from the previous sections that appreciation of validation over evaluation has, implicitly if not explicitly, increased in recent years. As Clunie et al. (2018) emphasised, we need more robust strategies that address the why, how, and when questions. There is a definite, pressing need for clarity in TEL validation.

7.3 METHODOLOGY

Prior to conducting our systematic review, we formulated a protocol conforming to the PRISMA-P checklist (Moher et al., 2015) and NIRO-SR guidelines (Topor et al., 2020). The protocol prescribed the steps of our systematic review and helped to ensure that our process was in accordance with the PRISMA guideline for reporting systematic reviews (Page et al., 2021). In this section, we will describe the core elements of our review methodology and deviations from the protocol. A more detailed description of our methodology can be found in the protocol, which we have made available in an open-source project along with our data. [1] All actions have been recorded with time stamps in our open-source project, for complete transparency. To our knowledge, this is the first time in the field of TEL that a systematic review protocol was made available open access prior to publication of the review.

1 https://osf.io/g2s56/

Table 7.1: Search terms and synonyms used in our database search. Terms must be included in the abstracts or titles of studies.

	SEARCH TERMS
	"technology-enhanced learning" OR "technology enhanced learning" OR "e-learning" OR "mobile learning" OR "digital learning" OR "electronic learning" OR "distance-learning" OR "web-based learning" OR "computer-based learning" OR "virtual learning"
AND	"validity" OR "validation" OR "quality" OR "evaluation"
AND	"criteria" OR "criterion" OR "dimension" OR "type" OR "aspect"

7.3.1 *Search strategy*

For our search we used the following databases: ACM Digital Library, IEEE Xplore, PubMed, and Web of Science. We included peer-reviewed journal and conference articles written in the English language. We consciously chose not to exclude any studies based on their publication date, since we aimed to analyse the use of validity criteria over time. The search terms used are listed in Table 7.1.

The search process produced 1,566 results, of which 1,256 remained after deduplication and removal of results that were not peer-reviewed journal or conference articles. The 1,256 publications served as input for our title and abstract screening phase, where we included studies that satisfied the following criteria:

1. Study design: the study reports on a TEL intervention in a real-world environment,
2. Participants: the study concerns a population of learners or educators,
3. Technology: the study discusses a technology with a direct impact on the learning experience,
4. Validity criteria: the study evaluates the TEL intervention using at least one clearly defined validity criterion.

The reason for focusing on intervention studies is that they are able to address validity criteria covering the full spectrum from design to implementation, hopefully preventing the introduction of a bias in validity criteria purely caused by the type of study considered. Additionally, by selecting only intervention studies, we establish a focused scope for this review. In future reviews, broadening this scope could offer valuable insights. A table summarising all inclusion and exclusion criteria can be found in our protocol.

7.3.2 *Selection process*

We used the ASReview screening software (van de Schoot et al., 2021) to perform title and abstract screening. ASReview optimises the title and abstract

screening process through the use of active learning, whereby reviewers are presented with the most relevant studies first. We initialised the ASReview process with two reviewers who independently screened a set of 100 randomly sampled articles. Both reviewers agreed on the exclusion of 87 articles and the inclusion of 7 articles. For 6 articles there was initial disagreement, leading to a Cohen's kappa of $\kappa = 0.67$. All disagreements resulted from different interpretations of abstracts, rather than from a fundamentally different understanding of which studies should be included. After discussion with two further reviewers, unanimous agreement was reached and the screening process was continued.

We estimated the total number of relevant articles in our set based on the random sample of 100 studies. We included 11 of the 100 studies. Our stopping criterion for the main screening phase specified that the reviewer should stop once they had reached 95% of the estimated number of relevant papers or had encountered 20 irrelevant articles in a row. The estimated number of relevant papers was $1,256 \times 0.11 = 138.16$, implying that the criterion was to stop screening once we had reached a total of 132 ($138.16 \times 0.95 = 131.25$, rounded up) relevant papers. This approach is based on the approach used in earlier research (van Haastrecht, Sarhan, Yigit Ozkan, et al., 2021). The 132 relevant records were reached after screening 374 records in total.

After title and abstract screening, we attempted to retrieve full-text articles for all potentially relevant records. At this stage, two articles were excluded as the main text was not written in English, and one article was excluded because a similar study by the same author was included. Ten articles could not initially be retrieved. We managed to contact the authors of seven of these articles, resulting in one additional full text inclusion. In total, 12 articles were excluded at this stage, resulting in 120 remaining inclusions.

We then performed backward snowballing to identify additional studies that may have been missed through the database search. We sorted all inclusions randomly before initialising the snowballing procedure. We stopped the backward snowballing phase once we had reviewed at least 10 inclusions fully and consequently reached a point where 100 references in a row were considered irrelevant. The snowballing phase was conducted by a single reviewer, with a second reviewer screening all papers that the first reviewer marked for inclusion. We evaluated a total of 543 references, resulting in a further 34 inclusions, on which the two reviewers reached full agreement.

Figure 7.1 summarises the selection process using a PRISMA flowchart. The final step of assessing data quality is discussed next. At this stage, 107 papers remained after we had excluded 47 of our 154 inclusions based on their full text content.

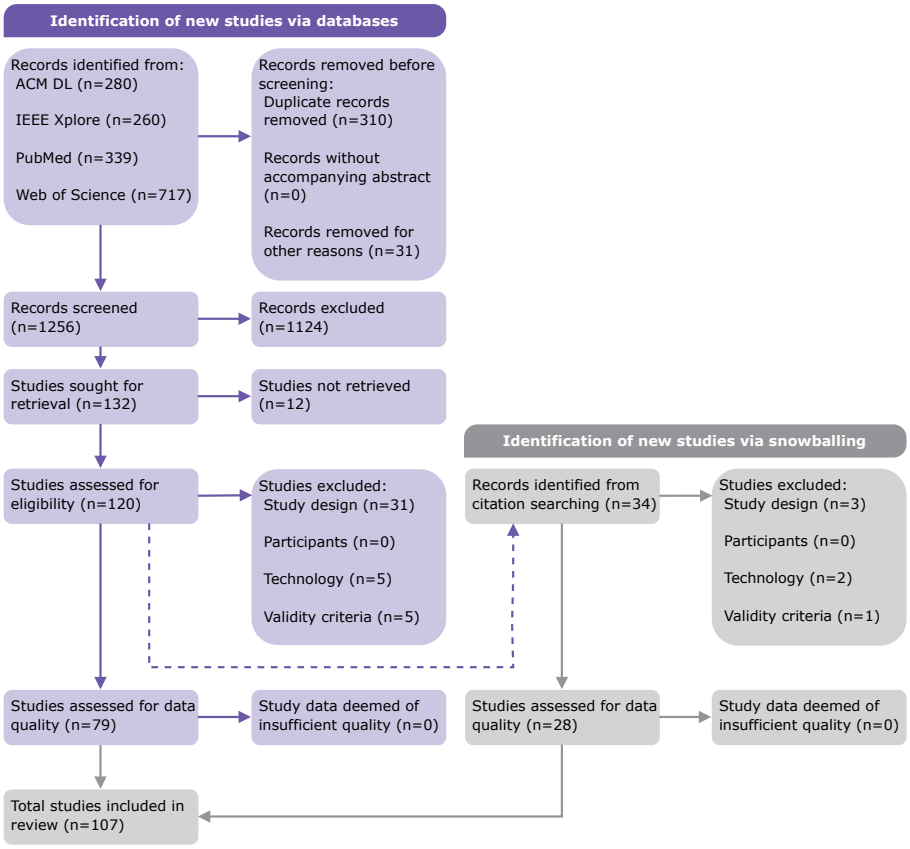


Figure 7.1: The PRISMA flowchart for our systematic review.

7.3.3 *Data extraction*

Data was extracted based on a data extraction form which can be found on our open data repository page. To ensure consistency, a pilot test of the data extraction form was conducted by two reviewers on a sample of 20 included studies. Any discrepancies were resolved through discussion. Our protocol initially specified that the pilot test would be conducted on a sample of 10 studies, but we found it necessary to have more data extraction examples to facilitate a detailed discussion, and therefore increased the sample to 20.

The data extraction form was revised before continuing with the full data extraction process, which was performed by a single reviewer. Compared to our original data extraction form as specified in our protocol, we included an item specifying the number of participants in a study and three items relating to the use of an evaluation or validation framework. We additionally noticed a focus on evaluation rather than validation in the first studies we analysed, and therefore decided to monitor the extent to which studies addressed the evaluation criteria as specified in the taxonomy of Lai and Bower (2019). Once data extraction was complete, two reviewers evaluated the completeness, accuracy, and consistency of the data. We did not encounter any issues, which is reflected in the fact that no studies were excluded due to insufficient data quality.

Table 7.2 lists the validity criteria that we consider in this systematic review. We provide possible synonyms for these criteria as observed in the literature (see e.g., Cronbach and Meehl (1955), Lincoln and Guba (1986), Mingers and Standing (2020), Straub (1989), and van Haastrecht, M. Brinkhuis, Peichl, et al. (2023)). Three changes were made compared to our protocol. Effectiveness replaced consequential validity as a main criterion, joyfulness was added as a criterion, and helpfulness was added as a synonym for usefulness. In cases where studies report validity criteria using different terminology or concepts, we attempted where possible to map these to the appropriate validity criteria listed in Table 7.2. We consciously chose not to prescriptively posit our own definitions for these validity criteria, but rather to take author's claims of assessing particular criteria at face value. If authors claim to assess trustworthiness, we included the criterion of trustworthiness for their study. To provide more insight into the implicit definitions used by researchers, we paired each included criterion with an accompanying quote from the relevant study, which can be found in the extracted data in our open access data repository.

For each validity criterion, we determined the research method used, the knowledge source for the evidence, and the outcome of how the criterion was assessed. For the research method item, we used the categories quantitative, qualitative, and mixed. If the evidence used to assess a validity criterion came from a qualitative approach such as an interview, the research method label for that criterion would be qualitative. From an epistemological point

Table 7.2: Validity criteria considered in our systematic review and their synonyms.

VALIDITY CRITERIA	SYNONYMS
Actionability	Practicability
Authenticity	Genuineness, originality, ecological validity
Confirmability	Auditability, accountability
Effectiveness	Consequential validity, impact, social validity
Consistency	-
Construct validity	Convergent validity, discriminant validity, specificity, structural validity
Content validity	Face validity, representativeness, comprehensiveness, objectivity [context: unbiased content]
Credibility	Authority
Criterion validity	Concurrent validity, predictive validity, empirical validity [context: predictive ability], accuracy
Dependability	-
Elegance	Appealingness, attractiveness, beauty, gracefulness
External validity	Generalisability, population validity, sample representativeness
Fairness	Impartiality, unbiasedness, equity
Internal validity	Causal validity
Joyfulness	Delightfulness
Meaningfulness	Significance [context: personal impact]
Parsimony	Simplicity
Relevance	Applicability, pertinence, suitability
Reliability	-
Replicability	Reproducibility, repeatability, objectivity [context: replicable research methodology]
Rigour	Thoroughness, soundness
Statistical validity	Statistical significance, empirical validity [context: correlation], statistical robustness
Transferability	Portability
Trustworthiness	Integrity
Understandability	Clarity, comprehensibility, interpretability, intuitiveness, transparancy
Usability	User-friendliness, accessibility, ease of use
Usefulness	Helpfulness, practicality, utility

of view, it is common to distinguish knowledge sourced from theoretical reasoning and knowledge sourced from practice. For the knowledge source item, we therefore used the categories theory and practice. If the evidence used to assess a validity criterion resulted from theoretical reasoning, as is often the case for statistical validity, the knowledge source label would be theory. Conversely, if the evidence resulted from feedback from students, the knowledge source label would be practice. Finally, we investigated whether the eventual outcome of validity criteria assessments was positive, negative, or mixed. If authors measured relevance and concluded based on their evidence that their solution was indeed relevant, the assessment was labelled as positive. A mixed assessment could be achieved if the evidence was inconclusive, or if certain evidence pointed to a positive assessment while other evidence pointed to a negative assessment.

7.4 RESULTS

Our main research question asks how we can characterise the landscape of validity criteria used in TEL research. To answer this question we formulated three sub-questions, which we will cover in the three subsections below.

7.4.1 *Which validity criteria are considered?*

Our first sub-question aimed to investigate which validity criteria TEL research considers and how researchers define and measure these criteria. Figure 7.2 shows how often each criterion was encountered in our inclusions. Effectiveness (82 appearances) and statistical validity (78) were the most commonly assessed criteria and Figure 7.2 shows that they tended to be positively assessed based on results from quantitative methods. In contrast, criteria such as external validity (34) and rigour (23) tended to be negatively assessed based on results from qualitative methods.

We extracted 298 criteria from our inclusions where the underlying argumentation resulted from a predominantly quantitative research method. Of these criteria, 34 (11.4%) were assessed negatively. Compare this to the 137 instances of predominantly qualitatively researched criteria, of which 77 (56.2%) were assessed negatively. Even when excluding external validity and rigour, which were often mentioned in the limitations section of research, 24 (30.0%) of the remaining 80 instances of predominantly qualitatively motivated criteria were assessed negatively. We can additionally observe that theoretically underpinned criteria were generally either argued for qualitatively and assessed negatively, or argued for quantitatively and assessed positively. This places these criteria at the extremes of the Figure 7.2 grid.

Figure 7.3 shows the research method used to assess each validity criterion instance. We excluded the two papers with more than 1,000 participants in

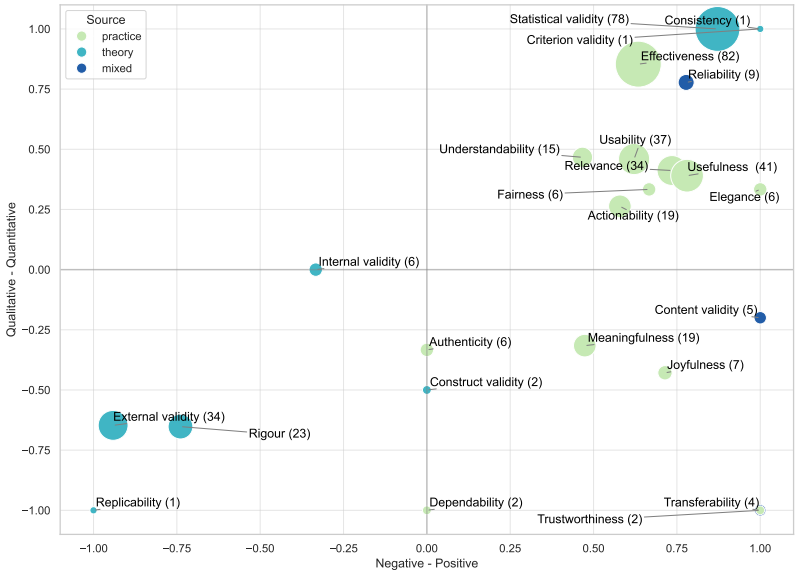


Figure 7.2: Bubble plot of validity criteria encountered in our inclusions, where each bubble is coloured depending on which knowledge source (on average) was used to assess it.

this plot to avoid readability issues due to scaling. When qualitative methods or mixed-methods were used by researchers, negative assessments were much more common than when quantitative methods were used. Since qualitative studies tend to have lower numbers of participants, one could wonder whether the connection between research method and assessment is caused by this mediating variable. To disentangle the number of participants and the research method, criteria are additionally visualised based on the number of participants in the study they were encountered in. The number of participants does not appear to be strongly correlated to the assessment.

Besides validity criteria, TEL researchers often consider constructs that are more directly tied to evaluation. With this in mind, we extracted data on the constructs used for TEL evaluation based on the list of constructs and construct themes presented in Lai and Bower (2019). Table 7.3 depicts the result of this work. Compared to Lai and Bower (2019), we observe relatively more criteria considered per paper, which can be explained by our confined search focusing only on studies that describe the criteria they use. We additionally observe less usage of established instruments and frameworks, which can partially be explained by the larger time window of our search and the fact that earlier papers used established instruments less frequently.

Figure 7.4 shows that differences with the results of Lai and Bower (2019) can be primarily attributed to a decrease in studies using established instru-

Table 7.3: Constructs used for TEL evaluation, following the exact structure of Table 6 in Lai and Bower (2019).

		Papers		Instruments	
THEMES/ASPECT (NO. OF PAPERS, %)	SUB-THEME CONSTRUCTS	NO.	%	ESTABLISHED	SELF-DEVELOPED
Learning (103, 96.3%)	Knowledge, achievement or performance	96	89.7%	25.0%	75.0%
	Cognitive load/effort (e.g., mental effort)	19	17.8%	36.8%	63.2%
	Skills development (e.g., interpersonal skills, motor skills, verbal and non verbal skills or communication skills)	39	36.4%	25.6%	74.4%
	Learning styles or learning strategies	27	25.2%	25.9%	74.1%
Affective Elements (82, 76.6%)	Perceptions, intentions or preferences	62	57.9%	24.2%	75.8%
	Engagement, motivation or enjoyment	50	46.7%	26.0%	74.0%
	Attitudes, values or beliefs	20	18.7%	30.0%	70.0%
	Emotional problems, anxiety or boredom	14	13.1%	35.7%	64.3%
	Self-efficacy	15	14.0%	20.0%	80.0%
Behavior (84, 78.5%)	Usage or participation	53	49.5%	20.8%	79.2%
	Interaction, collaboration or cooperation	52	48.6%	23.1%	76.9%
	Self-reflection, self-evaluation or self-regulation	20	18.7%	10.0%	90.0%
Design (59, 55.1%)	Course quality, course content, course structure, resources or overall design	59	55.1%	22.0%	78.0%
Technology (73, 68.2%)	Functionality	13	12.1%	38.5%	61.5%
	Perceived usefulness	45	42.1%	26.7%	73.3%
	Perceived ease of use	41	38.3%	24.4%	75.6%
	Adoption	3	2.8%	100.0%	0.0%
	Accessibility	34	31.8%	14.7%	85.3%
Teaching/Pedagogy (56, 52.3%)	Pedagogical practice, teaching strategies or teaching quality/credibility	49	45.8%	22.4%	77.6%
	Feedback	28	26.2%	32.1%	67.9%
Presence (10, 9.3%)	Social presence, co-presence or community	10	9.3%	20.0%	80.0%
	Presence in the environment	0	0.0%	-	-
Institutional Environment (11, 10.3%)	Institutional - institutional capacity, institutional intervention, institutional policy or institutional support	8	7.5%	25.0%	75.0%
	External environment/factors	3	2.8%	66.7%	33.3%

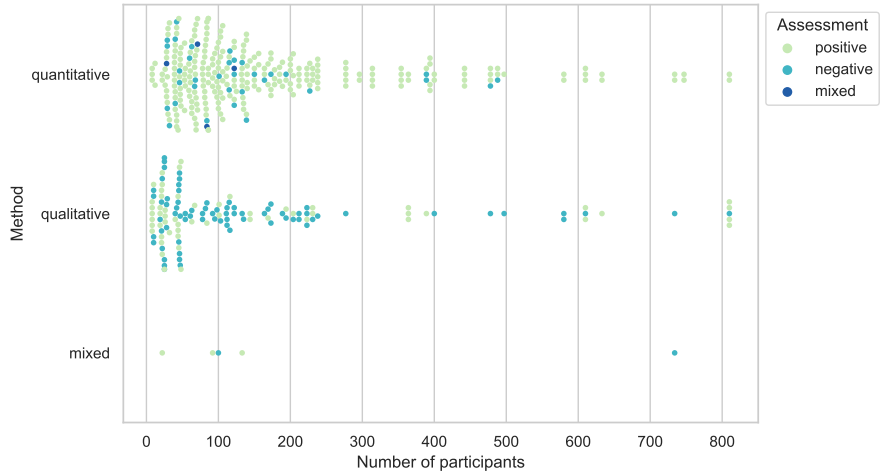


Figure 7.3: Plot of validity criteria occurrences, split by the number of participants in the corresponding study and the method used to assess the criteria. Criteria are coloured by whether the assessment was positive, negative, or mixed.

ments to evaluate learning and affective elements, and an increase in studies using self-developed instruments to evaluate technology and teaching/pedagogy. Altogether, 27 of our 107 inclusions used an established evaluation or validation framework, with the most commonly used framework employed only 4 times. This fragmentation in the use of frameworks was also found by Lai and Bower (2019), where the most common framework appeared 20 times among their 243 inclusions that applied an established instrument.

7.4.2 *How does criteria prevalence change over time?*

Our second sub-question asked how the prevalence of validity criteria has changed over time in TEL research. Figure 7.5 shows the percentage contributions of the ten most commonly encountered criteria in our inclusions over time. The bar plot stacks the top ten criteria from most frequently occurring overall at the bottom (effectiveness) to least frequently occurring at the top (understandability). The height of each individual bar within a year represents the percentage contribution of a criterion. The total height for a particular year represents the percentage contribution of the top ten criteria. We observe that the most frequently occurring criteria overall tend to be the most frequently occurring criteria per year. This is a first signal of the temporal stability of the validity criteria landscape in TEL.

Figure 7.6 is a variant of the Figure 7.2 bubble plot, but with each subplot showing criteria prevalence during the period of the subplot title. Time win-

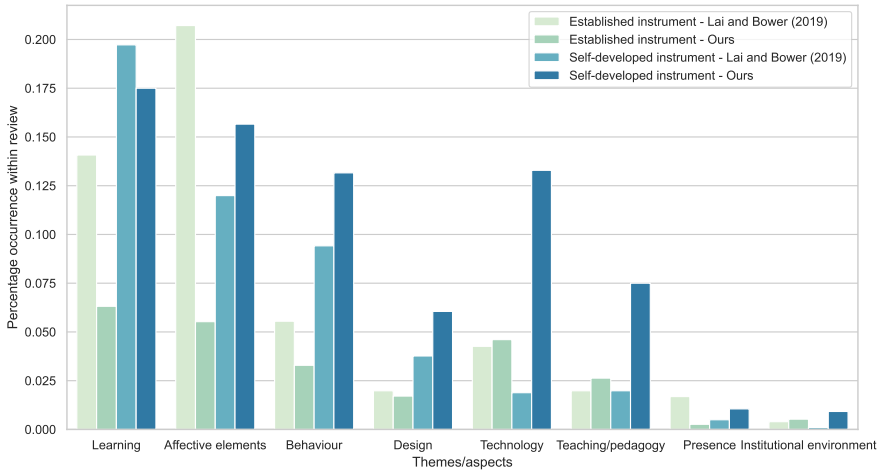


Figure 7.4: Comparison of the percentage occurrence of evaluation themes/aspects in Lai and Bower (2019) and this review.

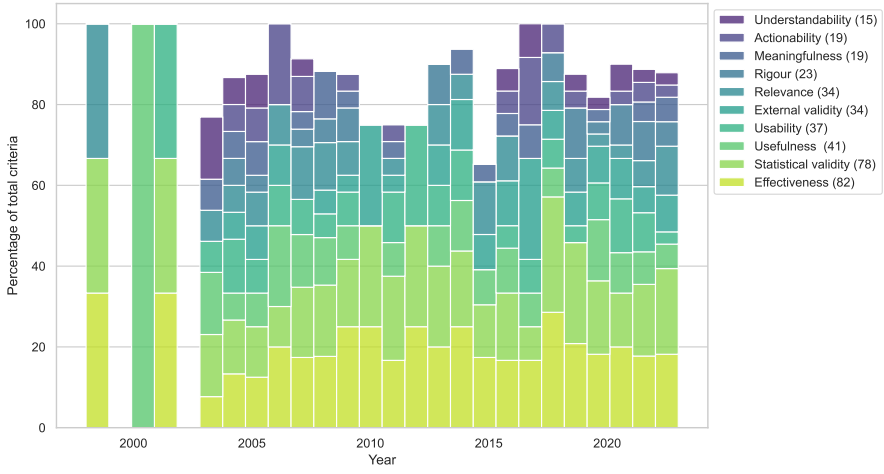


Figure 7.5: The ten most frequently encountered criteria overall, plotted per year from 1998-2023.

dows were chosen such that individual plots contain roughly equal numbers of validity criteria instances. Although we observe movement in the placement of certain criteria, the overall picture is relatively constant, with criteria that are coloured yellow and light green remaining in the top right and criteria that are coloured dark blue remaining in the bottom left. Of the 24 criteria shown in Figure 7.2, 20 are already included in the 1998-2009 plot. Of these 20, 17 are in the same quadrant overall as they were in 1998-2009, and 15 are both in the same quadrant and have the same knowledge source categorisation. The three criteria where the quadrant changes are elegance, fairness, and usefulness. In each case, the quadrant change was from bottom-right in 1998-2009 to top-right overall, meaning these positively assessed criteria were more commonly evaluated using quantitative methods in later years. Elegance and fairness had been assessed just once and twice, respectively, by 2009. The quadrant change for usefulness is more significant, as it had been assessed 15 times by 2009. However, the change from being primarily qualitatively assessed (y=-0.07) to being primarily quantitatively assessed (y=0.39) was not major. Figure 7.6 points to stable definitions and interpretations of criteria over time, thereby providing an affirmative answer to the question: Is there a common ground from which to build a comprehensive validation framework?

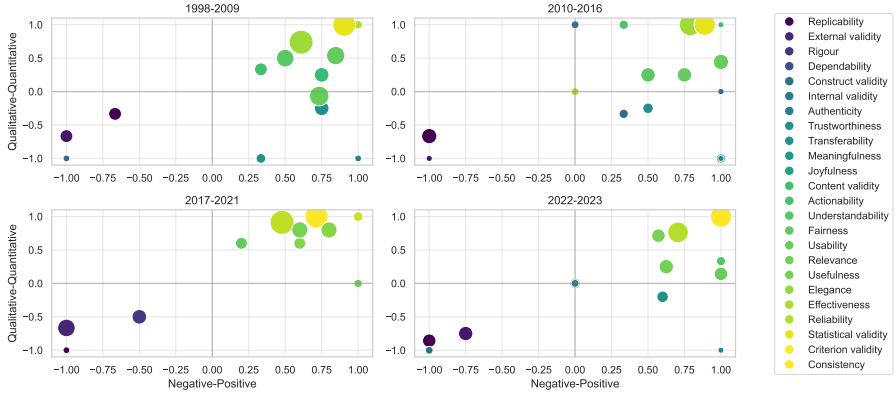


Figure 7.6: A grid of bubble plots visualising validity criteria positioning in different periods of time. Colours are determined by the position of a criterion in the overall plot of Figure 7.2.

7.4.3 *What epistemological patterns do we observe?*

Our final sub-question asked whether we observe any epistemological patterns in the connections between validity criteria. We concluded from Figure 7.2 and Figure 7.3 that there are observable relations between the method used to assess a criterion, the knowledge source used to inform this assessment,

and the eventual outcome of the assessment. However, these figures do not allow us to analyse the connections between validity criteria. Figure 7.7 represents a network visualisation of validity criteria, where the edge weights are determined by the relative co-occurrence of the target criterion in the papers where the source criterion was assessed. Node sizes are determined by how often a criterion was encountered and node colour is determined by whether a criterion was largely positively assessed (green) or largely negatively assessed (red). A reduced version of the total network is shown, as we only depict edges with co-occurrence scores of at least 90% of the maximum co-occurrence score per criterion. For example, the large edge weight for the edge going from dependability to authenticity indicates that the criterion of authenticity was encountered more often in the papers assessing dependability than we would expect based on the prevalence of authenticity as a criterion.

Figure 7.7 shows several clusters of validity criteria that are interconnected, as well as pairs of criteria such as fairness and transferability. The lack of strong connections emanating from statistical validity and effectiveness points to the ubiquity of these criteria. Even when effectiveness co-occurs with other criteria quite often, the corresponding edge weights will still be relatively small since the base probability of co-occurrence with effectiveness is high. Another explanation for the lack of strong connections could be that researchers tend to judge these criteria to be essential to their studies, regardless of the type of study. Metaphorically, effectiveness and statistical validity are acquainted to every criterion, but true friends with none.

One way Figure 7.7 can be useful is in helping to select criteria that together form an epistemologically complete set. When designing a validation framework for TEL, one might start with the inclusion of the top ten criteria shown in Figure 7.5. Figure 7.7 can then help to unearth which clusters of validity criteria would not yet be covered by this initial set, such as the pair internal validity-construct validity and the pair fairness-transferability. An extended framework that incorporates internal validity and fairness could then be considered epistemologically more comprehensive.

7.5 DISCUSSION

The results presented in the previous section provide answers to the research questions we posed, but also raise new questions that we will discuss further.

7.5.1 *A problematic hierarchy of validity criteria*

Figure 7.2, Figure 7.5, and Figure 7.6 visualise the prevalence and epistemological positioning of validity criteria. These visualisations suggest the existence of a hierarchy of validity criteria, which can be construed as problematic. At

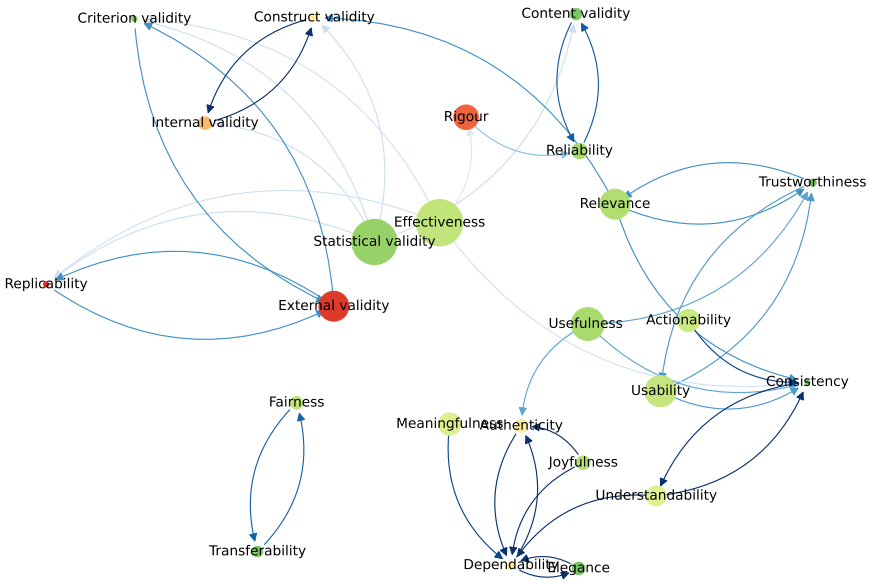


Figure 7.7: Network showing the relative co-occurrence of validity criteria.

the top of this hierarchy we find effectiveness and statistical validity. Around 75% of our inclusions assessed these criteria, and they were overwhelmingly assessed positively. However, Salehi et al. (2023) raise an important point regarding effectiveness and statistical validity that is generally not discussed in our inclusions. In a study of continuing professional development for 10,000 health workers in Ghana during the pandemic, the researchers mention in their discussion of e-learning effectiveness: "While these effect sizes are useful in painting an overall picture, with education evaluation, a 'small' effect size on a difficult-to-change variable (e.g., attitude toward recommending the vaccine) could be as valuable as a larger effect size on something easier to change (e.g., knowledge)" (Salehi et al., 2023, p. 10). Valuing particular criteria highly is not problematic in itself, but, as Salehi et al. (2023) point out, it is vital to critically contextualise validity evidence.

At the bottom of the hierarchy we find external validity, often termed generalisability, and rigour. External validity was assessed 34 times, with only one study reaching a positive conclusion. Of the 33 negative assessments, 28 times researchers mentioned that that their study focused on one educational context, and that this implies their results do not generalise to other contexts. Interestingly, the criterion of transferability, which can be seen as a counterpart to external validity (see, e.g. van Haastrecht, M. Brinkhuis, Peichl, et al. (2023)), was assessed positively 100% of the time. This points to the feasibility of

designing studies that produce generalisable results. An illustrative example is the one study that assessed external validity positively, which conducted a multi-centre randomised controlled trial (Vivekananda-Schmidt et al., 2005).

For studies that reached a negative conclusion about the rigour of their approach, common issues that were mentioned were the possibility of accidental exposure of the control group to the treatment (Tsai, 2010), the inability to mitigate certain biases due to the methodology used (Whitaker et al., 2007), and the overall lack of control over the experimental situation encountered during the COVID pandemic (Başağaoğlu Demirekin and Buyukcavus, 2022). Yet, there were positive examples too. One randomised controlled trial stated: "the major strength of this study is the robust methodology and adherence to protocol for each candidate once randomised" (Brewer et al., 2021, p. 5). A study applying a qualitative analysis of student reflections during the COVID pandemic concluded that "the strength of the study is that it provides quite a comprehensive picture of the students' experiences" (Wojniusz et al., 2022, p. 8).

Concerns have been voiced in earlier work about the troubling manner in which TEL research distinguishes between validity criteria in the upper echelons of the hierarchy, such as statistical validity, and criteria lower down, such as external validity (van Haastrecht, M. Brinkhuis, Peichl, et al., 2023). We are not calling for all studies to prioritise every validity criterion, as this is impossible. Yet, we need to ensure that as a field we do not structurally ignore certain criteria, while structurally prioritising, but not critically contextualising, other criteria.

7.5.2 *Framework foundations without structure*

Perhaps the most important finding from Section 7.4 is that there exists temporal stability in the usage and epistemological interpretation of TEL validity criteria. Figure 7.5 and Figure 7.6 convincingly showed that the same validity criteria have been prioritised by researchers throughout the last decades, and that the manner in which they have been assessed has remained remarkably constant. Naturally, one might conclude that the necessary foundations are present for a consensus validation framework.

However, we observed in Table 7.3 that usage of established frameworks is minimal. Additionally, similarly to Lai and Bower (2019), we found that there is a high degree of fragmentation in the use of frameworks. We suggested in Section 7.4.3 that our network analysis in Figure 7.7 could be a useful aid in selecting an epistemologically complete set of validity criteria for a validation framework. But a set of validity criteria is only the basis for a framework. A comprehensive framework requires a structure within which these validity criteria should be assessed and related to each other.

The argument-based approach to validation could offer the exact structure that TEL validation is currently lacking. In Section 7.2.1, we covered several

validation frameworks that have been proposed in recent years building on the argument-based approach (Huggins-Manley et al., 2022; van Haastrecht, M. J. S. Brinkhuis, Wools, et al., 2023; Zhai et al., 2021). Yet, we also highlighted that these frameworks are currently rarely employed in TEL research. There are criticisms regarding how these frameworks deal with TEL process data (Zumbo et al., 2023) and concerns whether they leave sufficient space for legitimately diverse arguments (Addey et al., 2020). Nevertheless, with the frameworks that already exist and the temporal stability present in TEL validation, there is clear promise for future holistic validation frameworks such as those based on the argument-based approach.

7.5.3 *Quantitative positivity: correlation or causation?*

We highlighted in Section 7.4.1 that there exists a correlation between the research method used to gather evidence regarding a validity criterion and the eventual assessment outcome. Not a single criterion in the quantitative half of the diagram in Figure 7.2 was on average assessed negatively. The question is whether there are any causal factors at play. Our research design was not suited to answer any causal questions regarding the relationship between research method and assessment outcome. However, we can present hypotheses that can be investigated in further research. Based on our discussion of a validity criteria hierarchy, one hypothesis is that the correlation is caused by publication bias. If predominantly quantitative criteria are considered more important than predominantly qualitative criteria, studies with negative assessments regarding quantitatively researched validity criteria would be less likely to get published than studies with negative assessments based on qualitative methods. One way to assess the hypothesis that a publication bias offers an explanation for the trend we observe in Figure 7.2, would be to survey TEL researchers. The researchers could be asked whether they consider research with negative quantitative results fit for publishing and whether they have experienced papers with negative quantitative results being rejected more often than papers of comparable quality with negative qualitative results.

Another hypothesis is that it is not the researchers, but rather the participants, that are causing the observed correlation. Quantitative approaches, such as questionnaires using Likert scales, condense constructs down to a numerical scale. In his seminal qualitative research work, Geertz (1973) delineates how qualitative methods are in search of meaning whereas quantitative methods are in search of law. One of our inclusions that applied qualitative methods was Rossiter et al. (2024), a study explaining the design and evaluation of a mobile learning resource for university students. A telling example of how qualitative methods can leave room for meaning over law comes from a student quote regarding the new resource's trustworthiness. The student explained: "I think I sort of trusted it a bit more because it felt like it was made by you for me as opposed to very general random videos that might be on

the subject area" (Rossiter et al., 2024, p. 119). A quantitative approach would not allow space for the meaning behind the student's positive assessment, and would likely abstract away this individual opinion into an aggregated number that serves as our law. A hypothesis to explain the correlation we observe in Figure 7.2 could thus be that quantitative methods leave less room for nuanced assessments and inadvertently hide negative or mixed feedback. A way to test this hypothesis would be to assess a set of constructs both quantitatively and qualitatively in a controlled environment. One could then examine whether quantitatively assessed constructs are evaluated significantly more positively.

7.5.4 *Limitations and threats to validity*

We should mention that this study has its limitations, along with potential threats to the validity of our conclusions. Firstly, although the search strategy we employed was geared at capturing all relevant studies for our systematic review, we cannot rule out the possibility that relevant papers were missed. An example of studies we may have missed are those that use wording in their title and abstract that deviate from the terminology of our search query. For example, we did not use the term 'online learning' in our original query. However, our systematic review process incorporating ASReview mitigates this risk by allowing for a broad database search with many related terms, and we additionally included a snowballing step which allowed us to identify relevant papers independently from our search query. For the case of the omitted term 'online learning,' our broad search strategy resulted in nevertheless having 22 of 107 papers including this term in the title or abstract. Additionally, only 4 of the 28 snowballing inclusions used the term 'online learning,' demonstrating that our search query did not miss disproportionately many studies for this term.

A potential threat to validity is the bias that the reviewers may have introduced into our screening process. Reviewers may have had personal biases that influenced which studies were included and how data was extracted. We believe the process we specified in our protocol and carried out for this study, where multiple reviewers were involved at each step of the systematic review, helped to minimise the risk associated with individual reviewer bias. Furthermore, by making our protocol available within an open-source project, we are transparent about our process and facilitate potential replication of our review.

Finally, although the 107 included papers and 440 extracted validity criteria constitute a comprehensive representation of the TEL literature, we have seen in Section 7.4 that certain criteria listed in Table 7.2 were either not encountered or rarely encountered. This could imply that our network analysis produced different results than if a larger set of papers would have been considered. The strictness with which reviewers followed our systematic

review protocol significantly decreases the probability that a replication study would find decidedly different results, but we would certainly welcome a large-scale systematic review that would enable deeper insights into the connections between TEL validity criteria.

7.6 CONCLUSION AND FUTURE WORK

Technological innovations have provided a diverse array of opportunities to optimise educational environments, but have also introduced new challenges in assessing the validity of novel solutions. We have seen in this chapter that the use of evaluation and validation frameworks in TEL research is rare and fragmented. However, we found that there is a clear light at the end of the validation tunnel. We demonstrated that the TEL validity criteria landscape has been remarkably stable over time. Both the types of validity criteria that are most commonly assessed and their epistemological positioning have stayed relatively constant over the past two decades. The stability in validity criteria usage and definitions offers a solid foundation from which to build future validation frameworks, where we highlighted the promise of argument-based validation to serve as the guiding structure.

There is a long road ahead before the use of holistic validation frameworks becomes commonplace in TEL research. Existing argument-based validation frameworks need to continually adapt to the changing world, with a constant need to recognise diverse perspectives and epistemologies. In our discussion section, we outlined several open questions whose answers would aid progress towards more holistic validation strategies. We observed a clear correlation between the research method used to assess validity criteria and the outcome of the assessment. Further research will need to determine whether the cause for this correlation lies with publication bias on the side of the research field, or with the inherent challenge of uncovering nuance and meaning using quantitative methods. Finally, future work will need to critically examine the problematic hierarchy of validity criteria that currently exists. We argue for a situation where validity criteria are prioritised critically based on their contextual relevance, rather than selected blindly based on their perceived importance.

8

VAST: VALIDATING SOCIO-TECHNICAL SYSTEMS

The influx of technology in education has made it increasingly difficult to assess the validity of educational assessments. The field of information systems often ignores the social dimension during validation, whereas educational research neglects the technical dimensions of designed instruments. The inseparability of social and technical elements forms the bedrock of socio-technical systems. Therefore, the current lack of validation approaches that address both dimensions is a significant gap. We address this gap by introducing VAST: a validation framework for e-assessment solutions. Examples of such solutions are technology-enhanced learning systems and e-health applications. Using multi-grounded action research as our methodology, we investigate how we can synthesise existing knowledge from information systems and educational measurement to construct our validation framework. We develop an extensive user guideline complementing our framework and find through expert interviews that VAST facilitates a comprehensive, practical approach to validating e-assessment solutions.

The contents of this chapter are based on: van Haastrecht, M. J. S. Brinkhuis, Wools, et al. (2023). VAST: a practical validation framework for e-assessment solutions. Information Systems and e-Business Management.

8.1 INTRODUCTION

Educational assessments have to clear various hurdles before being used in practice. The test of validity is recognised as the most indispensable of these hurdles. Naturally, this has led to a flourishing discussion on validity theory and validation frameworks in the educational field. Regarding traditional forms of assessment, we have reached a point in the debate where most of the dust has settled. However, the influx of technology in education has altered the playing field. Technology introduces new possibilities for assessments, such as evaluating collaborative problem-solving skills (Stadler et al., 2020) and using learner behaviour analytics (Douglas et al., 2020). Yet, electronic assessments (e-assessments) also pose new challenges for validation. Tests can now be more interactive and complex (Mislevy, 2016), threatening our ability to judge validity due to decreasing transparency (Wools, Molenaar, et al., 2019). There is a need for e-assessment validation frameworks and that need is currently not catered to by the two fields from which we might expect a contribution: information systems (IS) and educational measurement.

The use of technology poses new questions regarding the validity of our tests but also necessitates a validity assessment of the technology itself. There is consensus in the IS field that a comprehensive evaluation is crucial when designing new artefacts (Hevner et al., 2004; Peffers et al., 2007). Action design research even considers the development and evaluation of an artefact to be inseparable (Sein et al., 2011). Nevertheless, the "discussion of evaluation activities and methods" remains limited (Pries-Heje et al., 2008) and current frameworks commonly offer "little or no guidance" to researchers performing evaluations (Venable et al., 2016). An inclination towards formulating general frameworks is a potential cause of the lack of guidance. Criteria that "can be applied to all research approaches" (Mingers and Standing, 2020) point to a focus on generality rather than specificity.

In educational measurement, where validation has been a central topic for nearly a century, the problem of open-ended validation approaches was a motivator for Kane (1992) to formulate argument-based validation. Subsequent work has recognised the usability of Kane's framework, but concurrently identifies areas where it lacks practicability (Cook et al., 2015). To solve this issue, Hopster-den Otter et al. (2019) traded generality for practicability. They introduced a validation framework for formative assessment contexts which offers clear guidelines to practitioners on how to use the framework.

Validation of complex systems stands to gain the most from practical frameworks such as that of Hopster-den Otter et al. (2019). Not only is the burden of proof high for complex systems, but researchers struggle to collect sufficient validity evidence for these systems due to their uncontrolled nature (Broniatowski and C. Tucker, 2017). A clear and transparent process for validation is crucial in such a situation.

Socio-technical systems (STS) are recognised for their tendency towards complexity. In STS, complexity arises from the number of components and the interactions between those components. Yet, we lack validation frameworks for STS. IS validation targets instrument validation as the core pursuit (Straub, 1989), essentially ignoring the social dimension. This is surprising when we consider that some researchers state that "information systems are socio-technical systems" (van Aken, 2013). Conversely, educational measurement validation focuses on the interpretation and use of an assessment by a learner, but avoids judging the validity of the technology. In this chapter, we take a first step in addressing this issue.

Given the progress in developing practical validation frameworks for formative assessment, it is worth investigating whether we can apply these insights to validate STS projects. Specifically, we focus on socio-technical solutions with assessment as a central aim: e-assessment solutions. Stödberg (2012) defines e-assessment to entail any assessment making use of information and communication technologies, where "the entire assessment process, from designing assignments to storing the results" is included. Examples of e-assessment solutions are technology-enhanced learning systems (M. J. S. Brinkhuis et al., 2018), e-health applications (Eskes et al., 2016), and cybersecurity risk assessment applications (van Haastrecht, Sarhan, Shojaifar, et al., 2021). With the need for a comprehensive, practical validation approach for e-assessment solutions in mind, we formulate the following research question:

- **RQ**: How can e-assessment solutions be validated comprehensively and practically?

In the remainder of this chapter, we will first provide the background to this work in Section 8.2. Section 8.3 covers the research methodology we applied in answering our research questions. In Section 8.4, we introduce VAST: the first comprehensive validation framework for e-assessment solutions. Section 8.5 presents the results of our grounding procedure, which centred around applying our validation framework in the EU cybersecurity risk assessment project GEIGER (GEIGER Consortium, 2020). The feedback we received inspired the development of an extensive user guideline to accompany the VAST framework (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023). Where the VAST framework is the main theoretical contribution of this chapter, we envision the accompanying guideline to provide the most impact for practitioners striving to validate their solutions. We discuss the implications and limitations of our work in Section 8.6 and conclude in Section 8.7.

8.2 BACKGROUND

Thoughts on what constitutes validity have evolved over time and still differ across and within disciplines. In this section, we will cover those contributions

which help to understand the bigger picture of the validation literature, to create a common ground for the remainder of this chapter.

8.2.1 *Validation in educational measurement*

The field of educational measurement has close ties to psychological testing, which historically adopted a pluralist view on validity. Construct validity evolved from being an element in this pluralistic view to epitomising the overarching concept which unified all views on validity. A major proponent of this idea was Samuel Messick. Messick referred to the earlier pluralistic view as "fragmented and incomplete" and highlighted the need to integrate both "score meaning and social values in test interpretation and use" (Messick, 1995).

Although the validation framework Messick (1989) developed seemed to address many of the issues of earlier validation approaches, it was not very practical and "very open-ended" (Kane, 2013a). Kane (1992) introduced the argument-based approach to validation to address the open-ended nature of validation methods. Kane proposed a chain of inferences that form the interpretive argument, thereby giving guidance on "the kinds of inferences needed for the validation." Kane later extended this approach to an interpretation and use argument (IUA), aligning with the view of Messick that interpretation is not the only relevant dimension (Kane, 2004, 2013a).

Recent work has sought to provide guidelines on how to apply Kane's framework in particular contexts. Cook et al. (2015) provide a practical guide in the setting of medical education, noting that "Kane does not specify the order in which validity evidence should be collected and evaluated." Hopster-den Otter et al. (2019) extend the example inferences provided by Kane (2013b) for the context of formative assessment. The role of use is more prominent in formative assessment, which explains why Hopster-den Otter et al. (2019) chose to extend the IUA with additional use inferences. Although we have seen significant advances in the area of argument-based validation, Kane's framework has not yet been examined in assessment settings with a technological influx.

Modern times have seen the rise of technology-enhanced learning, with technology playing a part in our lives and education from an ever-younger age. With technology-enhanced learning becoming ubiquitous, one would expect an increased focus on validating e-assessments. Yet, although validating e-assessment requires specialised approaches (Wools, Molenaar, et al., 2019), no such approaches currently exist. This is a significant gap in the literature; a gap we aim to address in this work. To understand how we can best incorporate the technological viewpoint, we look towards the field that studies technological systems: information systems.

8.2.2 *Validation in information systems*

The seminal work of Straub (1989) on validity in IS outlines several validity types, as well as an order in which validation should address these types. Straub suggests to first conduct instrument validation, which consists of addressing content validity, construct validity, and reliability. Straub (1989) states that with content validity we answer the question: "Are instrument measures drawn from all possible measures of the properties under investigation?" This definition differs from the definition Cronbach and Meehl (1955) proposed in the educational measurement field, which states that test items should be an appropriate "sample of a universe in which the investigator is interested." Yet, the differences are somewhat superficial, as the underlying spirit is largely the same. Both definitions stress that content validity corresponds to how well we have sampled from the set of possible measurement items.

Straub's definition of construct validity also seems to depart from definitions as seen in Cronbach and Meehl (1955) and A. L. Brown and Campione (1996). Straub links construct validity to the question: "Do measures show stability across methodologies?" If stability is observed, we are dealing with valid constructs. Once more, however, the seeming disconnect with the more holistic definition of A. L. Brown and Campione (1996) is illusory. In later IS validation work based on Straub (1989), Mingers and Standing (2020) employ a definition which we feel strikes the right balance: "Do the measures converge on the construct and not on other distinct constructs?"

Reliability is the third element in Straub's instrument validation. Reliability answers the question of whether "measures show stability across the units of observation" (Straub, 1989). Although there is no direct analogue for this type of validity in the educational measurement field, inter-rater reliability is commonly incorporated in the inference chain of argument-based validation (Hopster-den Otter et al., 2019). Table 8.1 shows that Straub et al. (2004) mention Cohen's κ as a means of assessment for reliability. Cohen's κ is commonly used to measure inter-rater reliability.

Straub (1989) covers two further validity types: internal validity and statistical conclusion validity. Internal validity answers the question: "Are there untested rival hypotheses for the observed effects?" The underlying idea is that we should be confident in having identified the correct causal mechanisms at play in our setting. This is why we prefer to use the more direct definition of internal validity employed by Mingers and Standing (2020): "Are there alternative causal explanations for the observed data?" Statistical (conclusion) validity relates to the statistical robustness of validation results. If we can show that results are "unlikely to have occurred by chance" (Mingers and Standing, 2020), we add a further dimension to our overall validity claim.

Finally, Straub (1989) mentions the concept of external validity but states that "for the sake of brevity" it is not covered. In later work, Straub et al. (2004) link external validity to generalisability, but do not define the concept.

Table 8.1: Consolidated table of educational measurement and IS validity types considered in this chapter. Suggestions for means of assessment are provided for most validity types. Straub et al. (2004) and Mingers and Standing (2020) do not consider criterion validity and do not suggest means of assessment for internal and external validity.

TYPE	DEFINITION	MEANS OF ASSESSMENT
Construct validity	"Do the measures converge on the construct and not on other distinct constructs?" (Mingers and Standing, 2020)	Principal Component Analysis, Confirmatory Factor Analysis (Mingers and Standing, 2020; Straub et al., 2004)
Content validity	The extent to which measurement items are an appropriate sample from the universe of possible measurement items (Cronbach and Meehl, 1955; Straub, 1989).	Literature review, expert panel (Mingers and Standing, 2020; Straub et al., 2004)
Criterion validity	The extent to which test scores serving as an operationalisation of a construct correlate with an independent theoretical representation of the construct (i.e., the criterion) (Cronbach and Meehl, 1955).	Comparison to gold standard (Hopsterden Otter et al., 2019; Kane, 2013a)
External validity	"To what extent can the findings be generalised to other populations and settings?" (Mingers and Standing, 2020)	-
Internal validity	"Are there alternative causal explanations for the observed data?" (Mingers and Standing, 2020)	-
Reliability	"Do measures show stability across the units of observation?" (Straub, 1989)	Cronbach's α (Mingers and Standing, 2020; Straub et al., 2004), Cohen's κ (Straub et al., 2004)
Statistical validity	"Are the results sufficiently statistically robust that they are unlikely to have occurred by chance?" (Mingers and Standing, 2020)	R^2, F-test (Mingers and Standing, 2020), Structural Equation Modelling (Mingers and Standing, 2020; Straub et al., 2004)

Once more, we turn to the recent work of Mingers and Standing (2020) for our definition: "To what extent can the findings be generalised to other populations and settings?"

Criterion validity, a common concept in the educational measurement field, is largely ignored in the IS validation literature. We argue that in our context criterion validity is a vital element to consider alongside other validity types. This aligns with the prominent role Duolingo - the largest mobile language learning application - gives criterion validity in its validation approach. Duolingo's validity argument relies heavily on correlation with gold-standard language tests (Settles et al., 2020). Hence, we include criterion validity in our set of validity types presented in Table 8.1.

Since the work of Straub et al. (2004), the IS field has grown and changed considerably. The emergence of design science research saw the creation of new validation and evaluation frameworks. Work by Wieringa and Moralı (2012) and Venable et al. (2016) focused on suitable research methods for design science evaluation and validation. However, the initial focus Straub placed on instrument validity remained, meaning that the social element was still lacking in IS validation frameworks.

Frameworks linked to action research, such as that of Wieringa and Moralı (2012), more explicitly recognised the importance of the user. Yet, design science frameworks naturally target an evaluation of the designed artefact, rather than an assessment of validity. An example is the FEDS framework of Venable et al. (2016), which distinguishes the evaluation of purely technical

artefacts from the evaluation of artefacts involving a social component. This attention to social factors makes the framework more suited to STS, but an evaluation framework is not a validation framework. Where evaluation tends to focus on eliciting whether predefined performance indicators have been met, validation asks deeper questions on whether the designed artefact does what it was intended to do in its operational environment.

An additional problem is that current frameworks offer "little or no guidance" to researchers (Venable et al., 2016). The pluralistic view that is still dominant in IS validation today causes most frameworks to be complex and impractical. IS, like educational measurement, has not been able to solve the problem of open-ended validation. This is not a comforting thought when we consider that most IS validation frameworks do not recognise the social context of the instruments they are validating. We will require STS validation frameworks in the future and we need to avoid frameworks that are too general to be usable. Hence, we feel it is important to focus on the class of e-assessment solutions, where we can use insights from many decades of research in educational measurement validation to complement IS knowledge.

8.2.3 *Validation of e-assessment solutions*

In this section, we will cover three essential prerequisites for our validation framework: an existing validation framework to use as a basis, a modelling language to model e-assessment solutions, and an argumentation style for our argument-based validation approach. Regarding the first prerequisite, we use the Hopster-den Otter et al. (2019) formative assessment validation framework as the basis for our work. This framework extends the traditional IUA chain in argument-based validation with further inferences regarding use. The reasoning behind this extension is that a formative assessment validation framework must go beyond the inferences present in summative assessment frameworks. Formative assessment involves a translation of the outcome by the user to their situation, an evaluation of which actions they should take, and internalisation of the experience to learn.

Yet, the Hopster-den Otter et al. (2019) framework is not designed for STS. The terminology used (e.g., 'student learning') is specific to the classroom setting. To align the framework with STS, we draw on terminology from design research. Both educational and IS design research methods are employed when designing e-assessment solutions. Infusing the framework with terminology from these methods is our first step towards constructing an e-assessment validation framework. Figure 8.1 is the result of this process. The terms we introduce to the framework are inspired by the terminology used in the action design research work of Sein et al. (2011) and the educational design research work of McKenney and Reeves (2018).

Our second prerequisite is a modelling language to model the solution being validated. Any effort to validate an e-assessment solution must be

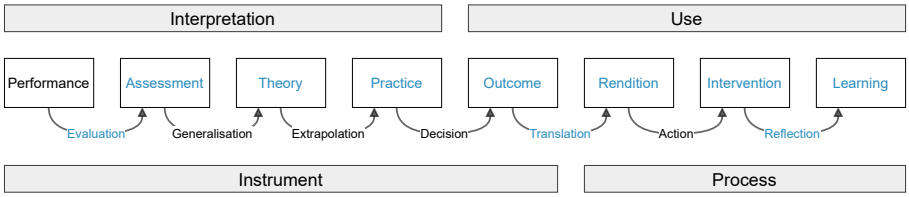


Figure 8.1: The inferences that make up the inference chain of the Hopster-den Otter et al. (2019) validation framework for formative assessment. Terminology that was adapted to suit our e-assessment setting is shown in blue.

predated by a description of that solution, consisting of the intended purpose and a representational model. We will assume that any researcher performing e-assessment validation has elicited functional and user requirements and is aware of the intended purpose of their system. This leaves the task of modelling the system.

Our STS model should, at minimum, include all relevant social and technical components and their interactions. If simplicity would not be a concern, flexible modelling languages such as Business Process Model and Notation (BPMN) and the Unified Modelling Language (UML) would be an ideal fit. However, BPMN and UML are notoriously complex modelling languages (Recker et al., 2009).

We should additionally acknowledge that we can treat the interpretation inferences of Figure 8.1 as being temporally independent, but that the same is not true for the use inferences. Use inferences depend on the thoughts and actions of users, which have a temporal structure. Hence, to address these inferences we must have a temporal model of our e-assessment solution. Finally, when evaluating use inferences it is preferable to initiate our argumentation from the user's perspective.

We have discerned that we require a modelling language that is not too complex, that allows for temporal dependencies, and that is user-oriented. We postulate that the answer lies in the use of user journey models. Any user journey representation that models all elements of an STS and their interactions satisfies the requirements we have put forth in this section. User journeys are temporal and user-oriented by nature. Therefore, a user journey modelling language that is not too complex can serve as the basis for our validation efforts. In this chapter, we employ the Customer Journey Modelling Language (CJML) (Halvorsrud et al., 2016; SINTEF Digital, 2022).

CJML models consist of temporally chained actions per actor. When an action constitutes an interaction with another actor in the system, CJML refers to this as a 'touchpoint.' Interactions have an initiator and a receiver. When multiple actors are involved, each actor has their own 'swimlane' in the CJML model. The corresponding diagram is termed a 'swimlane diagram.' The CJML swimlane diagram is the model we use in our validation framework.

Figure 8.2 shows an example swimlane diagram in the e-assessment setting. The user guideline (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023) that accompanies this chapter contains several examples detailing how to construct a CJML diagram.

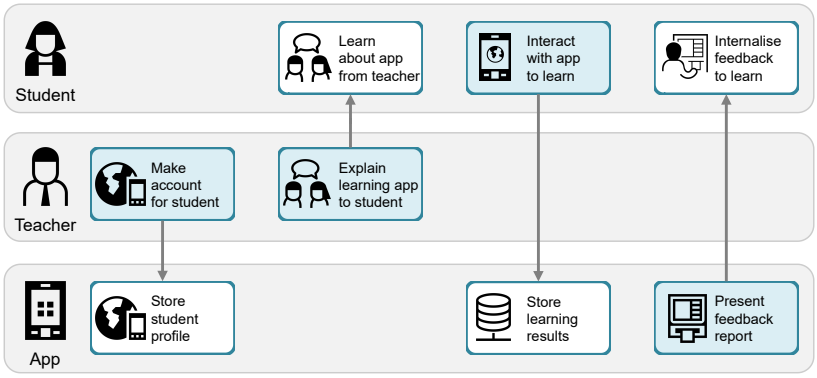


Figure 8.2: An example CJML swimlane diagram, where a student starts to use a mobile learning application. Each element of the system has a lane where actions are included in chronological order. When two elements of the system interact, the action of the actor initiating the interaction is coloured blue.

To address the final prerequisite for our validation framework, we will briefly cover the argumentation style we use within our argument-based validation approach. We choose to focus on Toulmin arguments since this style is commonly used in argument-based validation (Simon, 2008; Wools, Eggen, et al., 2010). Stephen Toulmin, a philosopher, introduced this structured style which divides argumentation into six components: claim, data, warrant, qualifier, rebuttal, and backing (Toulmin, 1958). Figure 8.3 depicts a Toulmin argument for the example of an online English language test.

8.3 METHODOLOGY

To synthesise theories from validation, modelling, and argumentation we require a flexible research methodology. We should build on existing theories and infuse our theory with insights from empirical work. Grounded theory is a research methodology suited to theory development. In its original definition, it was described as "the discovery of theory from data" (Glaser and A. L. Strauss, 1967). Grounded theory involves coding incidents found in the data into progressive abstractions to arrive at a theory, where 'incidents' are the basic units of analysis or ideas (Baskerville and Pries-Heje, 1999), and 'coding' involves the analysis and categorisation of incidents (Glaser and A. L. Strauss, 1967).

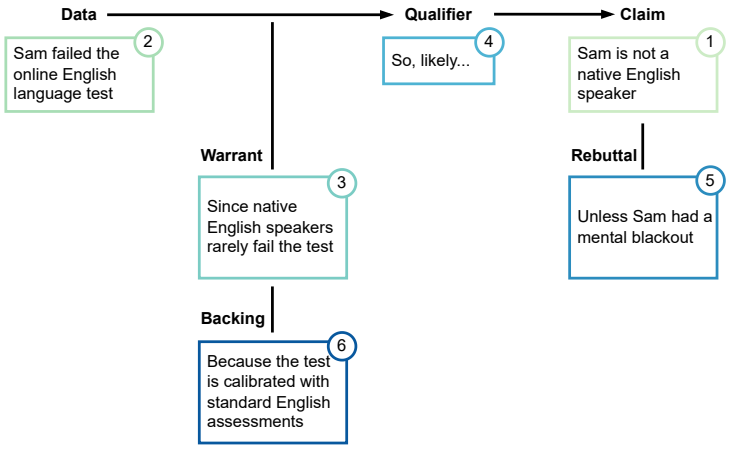


Figure 8.3: An example Toulmin argument for an online English language test. We want to make a claim (1) based on our data (2) and use a warrant (3) to support our claim. The qualifier (4) allows us to apply nuance to our claim. A rebuttal (5) can question the authority of our warrant, meaning we may require additional support to our warrant in the form of a backing (6).

Later extensions to grounded theory introduced three types of coding: open, axial, and selective (A. Strauss and Corbin, 1990). During open coding, the researcher aims to categorise essential incidents into concepts. Then, in axial coding, similar concepts are grouped into categories. Finally, selective coding works towards a core category, which from that point on is the main focus in the theorising process (Baskerville and Pries-Heje, 1999).

Grounded theory takes a purely inductive approach to theorising, meaning that in its strictest form grounded theory ignores established theories. The inductive approach has received heavy criticism, with some stating it constitutes a "loss of knowledge" (Goldkuhl and Cronholm, 2010). This led to the development of multi-grounded theory, where extant theories and knowledge receive a place in the theorising process. In multi-grounded theory, a researcher "constantly moves back and forward between data and preexisting knowledge or theories" (Thornberg, 2012).

Seeking to balance relevance-focused action research with rigour, Baskerville and Pries-Heje (1999) introduced the notion of grounded action research. The authors aimed for "a theory-rigorous and powerfully improved action research method," which remains practical and connected to organisational change (Baskerville and Pries-Heje, 1999). The multi-grounded variant of this approach soon emerged (Karlsson and Ågerfalk, 2007). Today, multi-grounded action research is positioned as the answer to how "knowledge development in action research [can] be clarified and improved" (Goldkuhl, Cronholm, and Lind, 2020). One way this manifests itself is in the three grounding approaches

present in multi-grounded action research: empirical grounding, theoretical grounding, and internal grounding. Emerging knowledge is grounded in empirical data through empirical grounding and in extant theories through theoretical grounding. Internal grounding helps to reflect on the emerging knowledge itself (Goldkuhl, Cronholm, and Lind, 2020). Figure 8.4 depicts the multi-grounded action research grounding procedure of our research. Extant theories contribute to the e-assessment validation framework through theoretical grounding and empirical data feeds into the emerging knowledge via empirical grounding. Lastly, expert evaluations provide internal grounding for our framework.

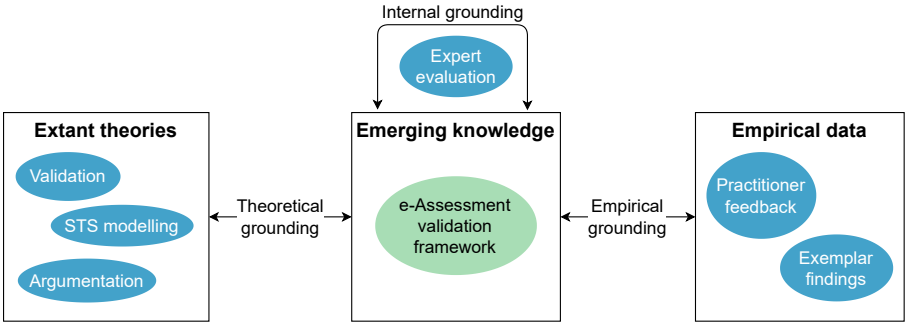


Figure 8.4: The grounding procedure of our multi-grounded action research methodology. Existing theories in validation, STS modelling, and argumentation provide theoretical grounding for our framework. We source empirical grounding from the practitioner feedback and exemplar findings that form our empirical data. Expert interviews help us to evaluate the internal cohesion of our emerging knowledge.

8.4 VAST

In this section, we propose VAST: an argument-based validation framework for e-assessment solutions. Traditional validation approaches consist of two main phases. First, a chain of claims specific to the project is constructed, which determines the inferences for which we need to provide arguments. Then, validity evidence is assembled to allow for a validity evaluation of our inference chain. However, in the complex setting of e-assessment, it is unclear where practitioners should source which evidence. VAST adds transparency to this process by inserting an additional step: modelling the system actions. The system model serves as a clarifying connector between the first and last steps in the validation process. Figure 8.5 presents the VAST framework. We will explain and motivate the three steps of VAST in the remainder of this section. For step-by-step instructions and practical examples of how to use the

VAST framework, we refer the reader to the VAST guideline (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023).

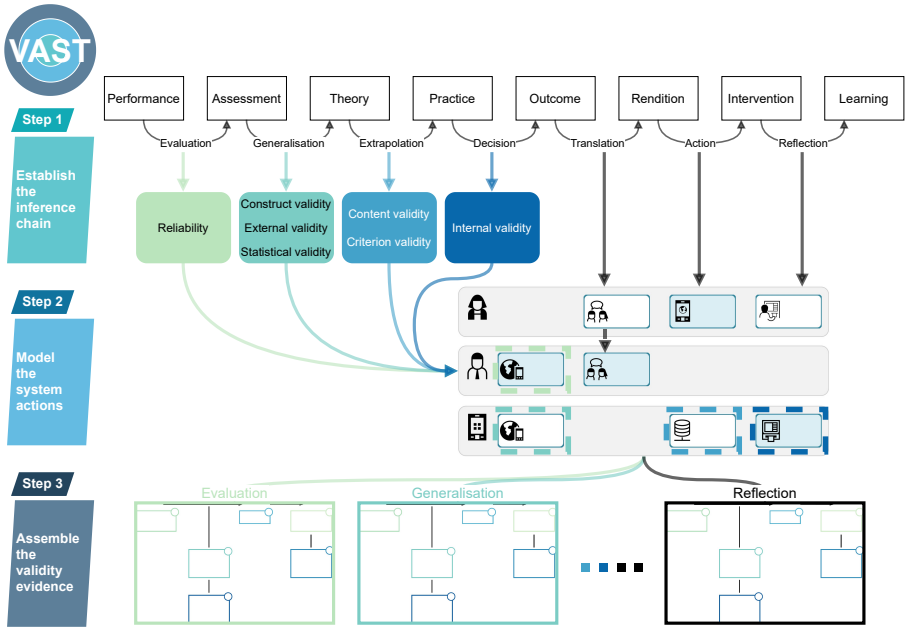


Figure 8.5: VAST: an argument-based validation framework for e-assessment solutions. VAST consists of three steps. Step 1 involves establishing the inference chain for the system being validated. By modelling the system actions in Step 2, we can match use inferences to user actions and instrument inferences to the remaining actions. Our model guides the assemblage of validity evidence in Step 3.

8.4.1 *Step 1: Establish the inference chain*

The first step within VAST consists of establishing the inference chain for the e-assessment solution at hand. We use our adapted version of the Hopster-den Otter et al. (2019) framework presented in Figure 8.1 as the starting point for this process. However, this is a general representation of an IUA chain, rather than a specific instance. Users of VAST will have to consider how the interpretation and use inferences materialise for their e-assessment solution. A vital prerequisite is that users have a clear idea of the objectives of their solution.

Part of this step will consist of making a first assessment of which inferences require more evidence than others. In certain systems, particular inferences will be redundant. As an example, consider the English language test we covered in Figure 8.3. If the test involves a diverse set of interactive written and

Table 8.2: Mapping of instrument inferences and validity types. We observe that the concepts related to instrument inferences relate to the concepts corresponding to particular validity types.

INFERENCE	CONCEPTS	VALIDITY TYPE	CONCEPTS
Evaluation	Consistency, inter-rater reliability	Reliability	Repeatability, stability
Generalisation	Theoretical constructs, different contexts, representative sample, control sampling error	Construct validity	Converge on construct
		External validity	Generalisation to other settings
		Statistical validity	Robust sample
Extrapolation	Accurate reflection of practice, theoretical tasks, compare to thorough assessments	Content validity	Appropriate sample, possible universe
		Criterion validity	Independent theoretical representation, comparison to gold standard
Decision	Underlying causal factors, outcome representation	Internal validity	Alternative causal explanations, observed data

oral exercises, the extrapolation inference taking us from theory to practice is largely obsolete. Although the option to prioritise inferences appears to introduce a layer of complexity to our framework, we want to stress that in principle all inferences should be considered. Only if a user of VAST is convinced that a particular inference is not relevant, should they disregard it.

Hopster-den Otter et al. (2019) connect their first four inferences to the instrument. In the following paragraphs, we will outline how we used our multi-grounded action research process to align the instrument inferences with IS validity theory. Table 8.2 shows the result of this work. In Section 8.4.2, we will investigate how we can synthesise the final three inferences with the e-assessment view.

In the evaluation inference, we assume that performance is consistently and reliably turned into assessment results. Inter-rater reliability is commonly mentioned as a possible source of evidence for this inference (Hopster-den Otter et al., 2019; Kane, 2013b). Mingers and Standing (2020) deem reliability to entail that results or responses are repeatable. This is similar to Straub (1989), who feels reliability should answer the question: "Do measures show stability across the units of observation?" We observe a clear connection between the concepts associated with the evaluation inference and with reliability. Hence, using the terminology of grounded theory, they are part of the same category.

In the generalisation inference, we assume the tasks of our assessment offer a sufficiently representative sample of the theoretical constructs we are aiming to represent (Hopster-den Otter et al., 2019). This ties the inference to our definition of construct validity outlined in Table 8.1. Additionally, it couples the inference to statistical validity. Statistical validity relates to whether our sampling approach is robust enough to rule out the possibility that results occurred by chance. This relates to the generalisation inference, which assumes that "tasks are sufficiently large to control sampling error" (Hopster-den Otter et al., 2019). We can observe from Table 8.2 that external validity relates to the generalisation inference. External validity addresses the following question: "To what extent can the findings be generalised to other

populations and settings?" (Mingers and Standing, 2020). This type of validity links to the generalisation inference, which extends the existing interpretation "to the expected performance over replications of the testing procedure (e.g., involving different test tasks, different testing contexts, different occasions, and raters)" (Kane, 2013a).

In the extrapolation inference, we assume that the theoretical tasks in the test domain accurately reflect practice (Hopster-den Otter et al., 2019). Content validity represents the extent to which test items are an appropriate "sample of a universe in which the investigator is interested" (Cronbach and Meehl, 1955). Content validity facilitates the extrapolation inference by motivating why our sample (theory) allows for an appropriate judgement regarding performance in the universe (practice) we are studying. A common way to support the extrapolation inference is to compare the results of our assessment to the results obtained by "assessments that cover the target domain more thoroughly" (Kane, 2013a). This corresponds to obtaining a gold standard result to compare to. This type of circular reasoning is both the link between criterion validity and the extrapolation inference and the "fundamental problem" (Kane, 2013a) of criterion validity.

The final inference we must account for is the decision inference, where a decision rule determines the outcome of our formative assessment. The choice of how to inform the user of the formative assessment outcome is vital, as it is the impetus for the formative process demarcated by the 'use' component of the IUA. This choice will be largely based on the causal factors that we assume to have generated the user's performance. With internal validity, we ask the question: "Are there alternative causal explanations for the observed data" (Mingers and Standing, 2020)? The internal validity of our e-assessment solution will determine whether we can formulate plausible backings for our decision inference. Hence, internal validity is the logical partner for the decision inference.

Our reasoning in the preceding paragraphs produced a coupling between the instrument inferences and the validity types of Table 8.1. The question remains how we can incorporate the inferences primarily related to use.

8.4.2 *Step 2: Model the system actions*

The second step in VAST consists of modelling the e-assessment system. We covered various STS modelling languages in Section 8.2.3, concluding that user journey modelling languages (specifically CJML) were best suited to our purpose. Figure 8.6 depicts the two stages involved in mapping the IUA inferences to our CJML model for the example covered in Figure 8.2. Recall that we are looking to inform the three use inferences: translation, action, and reflection. We posit that if any of the use inferences are of importance for an e-assessment solution, we can find a direct connection to at least one user

action corresponding to that inference. In our simple example of Figure 8.6, we see that each use inference connects to exactly one user action.

We connect the action of learning about the e-assessment application from a teacher to the translation inference. We reason that this introduction, whereby the teacher also learns from the student how they intend to use the application, will help in linking the eventual assessment to the student's circumstances. Given the inherent personal interactions that are present for the use inferences, we include the action of the teacher in this inference too. We denote this with a dotted, black arrow in Figure 8.6. If the user would have to perform additional actions themselves before the translation inference action, we would also connect these actions using dotted black arrows. Thus, we relate all actions to the translation inference that could directly or indirectly influence its interpretation in this context.

Similarly, we connect the action and reflection inferences to the CJML user actions. We connect the action inference to the interaction with the application and the reflection inference to the internalisation of feedback. Neither of these actions involves an interaction with another human actor. Rather, they constitute interactions with the application. Hence, we do not see any dotted arrows emanating from these actions.

Four actions remain unaccounted for. These are all the actions by actors that are not the student, except for those actions by human actors that involve direct interaction with the student. In a more general setting, we would refer to the student as the (main) social actor. Note that the actions that remain are not related to use, but rather to the instrument and preparatory work to enable later use. These are the actions that we can connect to the earlier inferences; the inferences regarding the interpretation and the instrument.

To couple the instrument inferences to the CJML diagram we can follow a more flexible approach. The instrument inferences do not need to abide by the temporal structure of the user journey model. Instead, we evaluate for each action which inference is most relevant. We circle the action using the colour of the most relevant inference. We see the result of this process in Figure 8.6. In our example, each inference corresponds to exactly one action. However, it is possible, and for larger e-assessment models often necessary, to map multiple actions to a single inference.

After completing the second step, the user of VAST will have gained further understanding of the system they are validating. Nevertheless, we have not yet assembled any validity evidence in the form of arguments. This is the focus of the third VAST step.

8.4.3 *Step 3: Assemble the validity evidence*

In the third and final step of VAST, the structured argumentation we discussed in Section 8.2.3 enters the stage. Figure 8.3 depicted the structure used in our arguments to motivate the inference from a datum to a claim. In the context of

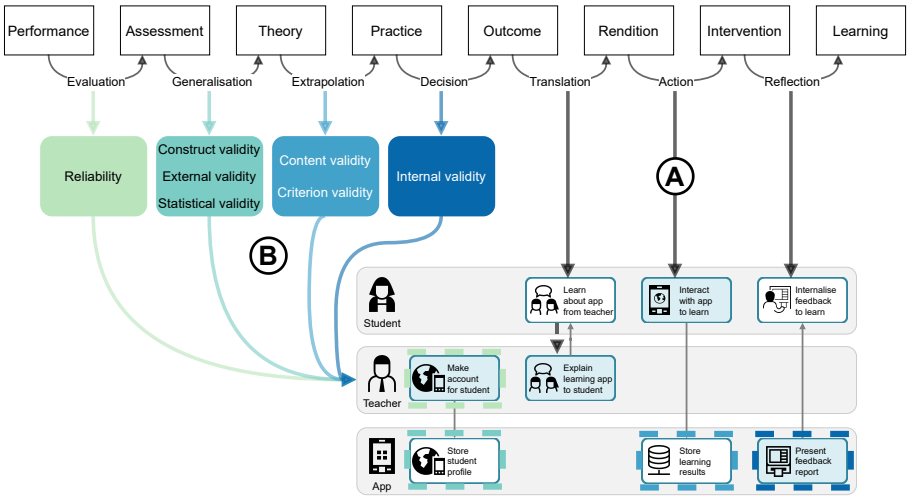


Figure 8.6: The two main stages involved in mapping the inference chain to the user journey model. First, we pair use inferences to user actions and interactions (A). The remaining actions are coupled to the inferences concerning the instrument (B).

our validation framework, we must provide argumentation for each inference, whereby the claim of the previous inference serves as the datum for the next inference.

Consider the evaluation inference, where we move from the performance datum to the assessment claim. We will at this stage have identified actions in our CJML diagram that relate to the evaluation inference and the corresponding validity type of reliability. Each action serves to inspire the relevant warrants, rebuttals, and backings that extend our argument. Although there is no absolute criterion to determine when an argument sufficiently motivates a claim, guidelines exist to assess the quality of argumentation. Erduran et al. (2004), for example, outline five levels of argumentation quality. From the lowest level 1 involving "a simple claim versus a counter-claim," we can improve to level 5 argumentation which "displays an extended argument with more than one rebuttal." Visually presenting the formulated arguments, as in the work of Wools, Eggen, et al. (2010), will then facilitate reviewers in assessing the quality of your argumentation.

Once we have provided sufficient evidence for the assessment claim, we proceed to the generalisation inference which connects the assessment datum to the theory claim. We continue along our inference chain until we have addressed all of our inferences. In this sense, the IUA and its argumentation serve "to specify what is being claimed" (Kane, 2013a). The final task is to assess the overall IUA with a validity argument, which "evaluates the plausibility of the proposed interpretations and uses" (Kane, 2013a). Kane

intends this to mean that the IUA is complete, coherent, and "supported by adequate evidence." VAST is structured to optimally address the validity argument.

Figure 8.5 depicts the three steps of establishing the inference chain, modelling the e-assessment solution, and assembling the validity evidence. Additionally, Figure 8.5 shows how the steps are connected to guide the user through the process. We believe the guidance provided within our framework allows us to counter the open-ended nature of validation and provide an actionable path towards validation. Although we have entrenched our framework in extant theories to provide theoretical grounding, we have not addressed the equally vital empirical and internal grounding within the multi-grounded action research grounding procedure. In Section 8.5, we turn our attention to practice to cover these further grounding procedures.

8.5 EVALUATING VAST

For our empirical and internal grounding, we applied VAST within the EU Horizon 2020 project GEIGER (GEIGER Consortium, 2020). GEIGER developed a cybersecurity risk assessment application for small businesses. The application helps raise employee awareness of cybersecurity threats and increase cybersecurity resilience. GEIGER assesses the cybersecurity risk faced by users and uses the outcome to offer personalised recommendations (van Haastrecht, Sarhan, Shojaifar, et al., 2021). By taking a formative approach to cybersecurity risk assessment, the GEIGER application forms an instance of the (formative) e-assessment solutions we are studying in this chapter.

To empirically ground VAST, we used an early variant of the framework to validate the GEIGER project. The details of this process are described in van Haastrecht, Spruit, et al. (2021). During six months of preparatory work, we gathered feedback on the first version of VAST from 13 different stakeholders across 14 sessions. We received comments that the framework did not offer enough practical guidance for validation. This feedback led us to include the second step of VAST, where the system is represented by a user journey model. The modelling step helped practitioners to connect abstract validation concepts to concrete user actions.

The updated version of our framework was further refined based on our first validation activities. These activities included an expert evaluation of the GEIGER content involving 14 stakeholders and user experience testing with our five use case partners (van Haastrecht, Spruit, et al., 2021). The findings from our practical application helped us to refine the step-wise approach of VAST, as it highlighted the necessity of forming a prioritisation among different validation activities. The refined variant of VAST was then further evaluated through interviews with validation experts.

To internally ground our framework, we interviewed three validation experts. We interviewed a senior researcher (SR) within the GEIGER project, an

Table 8.3: The current role, sector, and validation experience of our three interviewees. We use the ID to refer to the interviewees within the text.

ID	ROLE	SECTOR	VALIDATION EXPERIENCE
EA	External advisor	Private	15-20 years
PO	Project officer	Government	12 years
SR	Senior researcher	Academia	10 years

external advisor (EA) who is a member of the advisory board of GEIGER, and a European Commission representative with experience as a project officer (PO). All experts had at least ten years of validation experience at the time of the interview. Table 8.3 lists the details of the interviewees. The interviews consisted of a short introduction presentation explaining the GEIGER project and the VAST framework, followed by eight questions aimed at informing our internal grounding procedure. We list the interview questions in Table 8.4. Note that at the time of the interviews we had not yet developed the VAST guideline to accompany our framework.

With our main research question in mind, we asked the experts how VAST compared to traditional validation approaches regarding comprehensiveness and practicability. To make the concept of comprehensiveness more tractable for interviewees, we stated that this corresponds to coherence and completeness, using the terminology of Kane (2013a). EA and PO indicated that VAST would result in a much more coherent and complete validation process. SR stated that they could not compare VAST to earlier approaches in this way, since earlier approaches were always tailored to a specific project. Regarding practicability, EA and PO conveyed that VAST has the potential to at least be equally practical, given that users of the framework are well-prepared. SR suggested that more testing would be necessary to determine the practicability of VAST, although they too indicated that VAST has potential if it is supplemented with guidelines and practical examples on how to apply it.

Finally, EA and PO stated that they would likely recommend the use of VAST if they were to be involved in a future project of a similar nature. SR could imagine that they would recommend VAST given that adequate documentation and a practical, simple example of a VAST application exists. With internal grounding, we intend to investigate the "internal cohesion of the knowledge" being developed (Goldkuhl, 2004). Given the answers of our interviewees, VAST certainly exhibits internal cohesion. Nevertheless, there are areas for improvement, which we will cover in the following section.

8.6 DISCUSSION

Our grounding procedure demonstrated that although VAST helps to address the open-ended nature of validation, it cannot be considered a vali-

Table 8.4: Questions asked during the interviews with validation experts, in chronological order.

QUESTION	TYPE	OPTIONS
Please describe your previous validation experience (information systems, education, or other). How many years of experience do you have in your current role?	Open	-
What does validity constitute in your eyes? And validation?	Open	-
How do you view the original validation approach envisioned for GEIGER? Is it a similar approach to what you have encountered before?	Open	-
How do you view the VAST validation approach that was used for GEIGER? How appropriate do you think it is to build VAST on the argument-based formative assessment validation framework of Hopster-den Otter et al. (2019)?	Open	-
If you were to compare the VAST approach to the originally envisioned validation approach, how would you rate VAST in terms of coherence and completeness?	Likert	Much less, less, equal, more, much more
If you were to compare the VAST approach to the originally envisioned validation approach, how would you rate VAST in terms of practicability?	Likert	Much less, less, equal, more, much more
How likely are you to recommend the use of VAST for validation if you were to be involved in a future project of a similar nature?	Likert	Extremely unlikely, unlikely, neutral, likely, extremely likely
Is there anything else you would like to add? For example, something that you think can be improved in VAST.	Open	-

dation panacea. We have yet to see how VAST fares when applied to other e-assessment contexts, which themselves constitute only a fraction of all socio-technical systems. As we look to generalise, it is worth considering the observations of Addey et al. (2020). Though not outright disagreeing with the underlying push for clarity in Kane's argument-based validation, they observe that "in the quest for clarity and consensus, validity theory can become rarefied and idealised, and recognition of diversity diminished." Addey et al. (2020) note that Toulmin, who Kane builds on, shifted from an absolutist view on argumentation towards a more pluralistic one. Interestingly, this is in line with the view on validation we encounter in IS.

As we look to apply VAST in future work and generalise it to further socio-technical domains, we must always be wary of an overemphasis on clarity. We argued in our introduction that validation is inherently open-ended. When we take the pragmatic view of Kane too far, clarity becomes a requirement for successful validation, rather than a luxury. When this happens solutionism is just around the corner, especially in areas such as education where it already makes a regular appearance (McKenney and Reeves, 2021). Nevertheless, the reality of today's world is that complex systems exist and are continually being developed. As socio-technical systems increasingly become a part of our daily lives, we should not shy away from debating their validity. We believe frameworks such as VAST have a role to play in validation, as we strive towards clarity while recognising complexity.

Table 8.5 summarises the feedback we received in our expert interviews and the remarks of Addey et al. (2020) in three main suggestions for improvement of VAST. The first is to provide clarity where possible and appropriate. The experts we interviewed indicated that VAST would benefit from clear guidelines and supporting documentation, including practical examples of how to apply the framework. Focus groups could help us to improve the supporting

Table 8.5: The three axes of improvement identified for the VAST framework, resulting from a synthesis of the feedback from all interviewees. We briefly explain each concept and propose a possible research method we could use to investigate the implementation of each improvement.

IMPROVEMENT	EXPLANATION	RESEARCH METHOD
Clarify	Provide several practical examples on how to apply VAST, along with a clear step-by-step guideline and supporting documentation.	Focus groups
Modularise	Expand the scope of VAST outside of the e-assessment setting by providing custom inference chains for other STS classes.	Case studies
Visualise	Ensure that diverse perspectives on validity are recognised by designing a supporting tool where validity evidence can be assembled, debated, and visualised.	Educational design research

material in a collaborative, iterative fashion. To take a first step in addressing this axis of improvement and to signal our commitment to improving VAST, we created an extensive VAST guideline with practical examples to support future users (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023). We plan to use focus groups to help us in iteratively refining the VAST guideline.

The second suggestion is to transform VAST to a more modular approach. By providing custom inference chains for other STS classes, we can expand the scope of VAST. A series of case studies could help to determine which STS classes, and corresponding inference chains, could be validated with a more flexible variant of VAST.

Finally, to ensure that VAST does not contribute to a diminishing recognition of diversity, we should develop a supporting tool which promotes a lively debate on validity. We agree with Addey et al. (2020) that argument-based validation needs "a democratic space in which legitimately diverse arguments and intentions can be recognised, considered, assembled and displayed." Following a design research methodology such as educational design research could be an appropriate approach to create such a tool.

8.7 CONCLUSION

Socio-technical systems are complex and difficult to validate, meaning we often have to rely on validity assessments that address only parts of the system. We investigated how e-assessment solutions, a particular class of socio-technical systems, can be validated comprehensively and practically. We compared and synthesised ideas regarding validation from the educational measurement and information systems fields. This resulted in an adaptation of the Hopster-den Otter et al. (2019) validation framework to suit the context of e-assessment.

We then used a multi-grounded action research approach to aid the development of VAST: an argument-based validation framework for e-assessment solutions. VAST is the first validation framework that explicitly combines validity theory from educational measurement and information systems. VAST

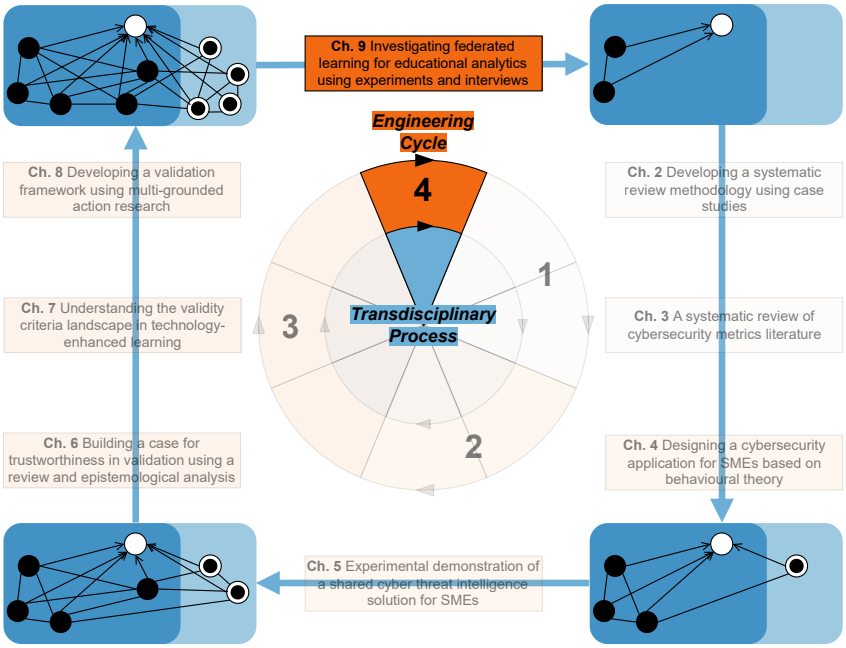
thereby addresses a significant gap that existed in the literature on socio-technical systems, namely the lack of validation approaches addressing both social and technical elements of the system being validated. We achieved this synthesis by identifying the commonalities between educational measurement inferences and information systems validity types.

Besides theoretical grounding, VAST resulted from empirical and internal grounding sourced from a practical implementation in the GEIGER project. We identified a need for clarity in the validation process, which VAST addresses by connecting inferences to concrete actions within the system. VAST additionally allows for transparent reporting of validation results by assembling validity evidence in the structure of Toulmin argumentation.

The validation experts we interviewed were assured of VAST's ability to facilitate a comprehensive and practical validation process. Still, the interviewees also provided suggestions for how to improve VAST. In future work, we hope to further VAST along the three axes of improvement identified by the experts: clarification, modularisation, and visualisation. We have already taken a first step in the area of clarification through the creation of an extensive VAST guideline containing practical examples (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023). We expect this guideline, which incorporates concrete example use cases, to be of value to both researchers and practitioners. The foundation VAST provides spurs our confidence about the future of holistic socio-technical systems validation.

Part IV

TREATMENT IMPLEMENTATION



9

FEDERATED LEARNING FOR EDUCATIONAL ANALYTICS

Concerns surrounding privacy and data protection are a primary contributor to the hesitation of institutions to adopt new educational technologies. Addressing these concerns could open the door to accelerated impact, but current state-of-the-art approaches centred around machine learning are heavily dependent on (personal) data. Privacy-preserving machine learning, in the form of federated learning, could offer a solution. However, federated learning has not been investigated in-depth within the context of educational analytics, and it is therefore unclear what its impact on model performance is. In this chapter, we compare performance across three different machine learning architectures (local learning, federated learning, and central learning) for three distinct prediction use cases (learning outcome, question correctness, and dropout). We find that federated learning consistently achieves comparable performance to central learning, but also that local learning remains competitive up to 20 local clients. We introduce FLAME, a novel metric that assists policymakers in their assessment of the privacy-performance trade-off, and conclude by discussing preliminary findings from a series of interviews with stakeholders we are conducting to unearth their views on federated learning for education.

The contents of this chapter are based on: van Haastrecht, M. Brinkhuis, and Spruit (2024). Federated Learning Analytics: Investigating the Privacy-Performance Trade-off in Machine Learning for Educational Analytics. Accepted at AIED 2024, prior to inclusion of interview materials in discussion.

9.1 INTRODUCTION

Driven by the promise of analytics to enable learning environment optimisation, education is now more datafied than ever (Williamson et al., 2020). The large-scale collection of learner data raises concerns regarding ethics, privacy, fairness, and trustworthiness (Gardner et al., 2023; van Haastrecht, M. Brinkhuis, Peichl, et al., 2023). Research tends to focus on the data protection measures educational institutions should implement to convince learners that they can be trusted as data fiduciaries (Jones et al., 2020). Examples of suggested measures concerning data that has already been collected are limiting the boundaries of access to student data, pseudonymisation and anonymisation of learner records, and using automated bias mitigation. However, approaches that assume that personal data has already been collected fail to address a fundamental question: Did we have to collect the data in the first place?

It is not trivial to motivate which, if any, educational optimisations would warrant an intrusion of student privacy. Institutes that hold student privacy in high regard may be of the opinion that collecting personal learning data is never warranted (Rubel and Jones, 2016). This puts educational analytics research in an uncomfortable position, as methods and applications commonly rely heavily on personal data. Machine learning models such as deep neural networks predicting learning outcomes (Waheed et al., 2020) and transformers facilitating student knowledge tracing (D. Shin et al., 2021) are deeply dependent on the availability of large amounts of data. On the surface, it seems as though these data-hungry machine learning models are incompatible with a policy of preserving student privacy. However, in recent years we have seen the development of machine learning architectures that promise the performance of machine learning without the threats to privacy posed by institute access to personal data.

Privacy-preserving machine learning architectures such as federated learning (McMahan et al., 2017), where only model parameters are shared with a centrally coordinating party, offer a promising future direction for educational analytics. Along with local learning, where nothing is shared, and central learning, where everything is shared, federated learning is among the major machine learning architectures to consider from a privacy perspective. We have recently seen the first studies investigating the promise of federated learning for educational analytics (Fachola et al., 2023; Guo and Zeng, 2020). However, to our knowledge, no study has systematically compared local learning, federated learning, and central learning across different datasets and use cases. This is a significant gap in the literature when we consider that privacy-preserving techniques could be the key to giving control back to students (Ekuban and Domingue, 2023).

In this chapter, we hope to take a first step in systematically investigating the promise of federated learning for learning analytics, which we term 'federated

learning analytics'. We compare the performance of local learning, federated learning, and central learning across three distinct use cases: learning outcome prediction, question correctness prediction, and dropout prediction. Our methodology is geared at answering our main research question:

- **RQ**: How does the privacy-performance trade-off for machine learning algorithms manifest itself in different educational analytics use cases?

9.2 BACKGROUND

Preserving the privacy of learners while actively collecting their data has long been recognised as a major challenge. It is evident that students should never be considered simply as sources of data, but rather as collaborators whose learning and development we are trying to serve (Slade and Prinsloo, 2013). However, although the importance of formulating and employing ethical and privacy principles was recognised early on, privacy concerns regularly played second fiddle due to the "enthusiasm for the possibilities offered by learning analytics" (Prinsloo and Slade, 2015). New legislation surrounding data protection introduced new perspectives. Besides ethical and privacy concerns, legal concerns began to drive decisions made at educational institutions. In the educational privacy framework DELICATE (Drachsler and Greller, 2016), the section on legitimacy contains the question: "Which data sources do you have already, and are they not enough?" Questions like these represented a major change of mindset. Researchers and practitioners recognised that collecting particular types of data is never warranted, and that "learning analytics is justifiable just to the extent that it does indeed promote autonomy" (Rubel and Jones, 2016).

Basic organisational and technical controls can help to preserve student privacy, but it is questionable whether this is sufficient to gain students' trust. Prinsloo and Slade (2015) convincingly argue that "the power to harvest, analyse and exploit data lies completely with the provider," rather than the student. The authors outline the importance of transparency towards students and of giving students the possibility to access and update their own information. However, the issue with these measures is that they still require the student to entrust multiple stakeholders with their personal data, keeping alive the privacy power imbalance between the student and the data fiduciary.

Levelling out the power balance is exactly what decentralised approaches have attempted to do in recent years, by enabling the sharing of student data in a way that can enhance both privacy and security within educational systems. Students thus regain some ownership over their data, helping to restore the power balance. Yet, using a decentralised architecture also introduces new challenges. The most prominent of these is how to maintain performant algorithms when not all data is available in one central data store. A study of several anonymisation and differential privacy techniques found that in

a GPA prediction task accuracy could drop from 76% to anywhere between 45-63% (Gursoy et al., 2017). Novel methods such as deep learning and transformers are notorious for requiring immense datasets to tune their parameters. How can we continue using these successful machine learning architectures when we do not have the data they so desperately need in one central location?

McMahan et al. (2017) introduced the concept of federated learning, where learning occurs over a federation of users referred to as clients. Rather than having to share data and parameters, clients train their model on local data and only share the parameter values of their model with the coordinating server. By averaging the parameters of all local clients, the resulting global model obtains better performance than if all local clients operated independently. Figure 9.1 visualises the scenarios of local learning, federated learning, and central learning. A fourth scenario was recently proposed where data is kept locally and parameters are not shared with a centrally coordinating server, but rather with other trusted parties via blockchain (Warnat-Herresthal et al., 2021). This architecture, termed swarm learning, is worth considering for educational institutions. However, we will not investigate it in detail within this chapter.

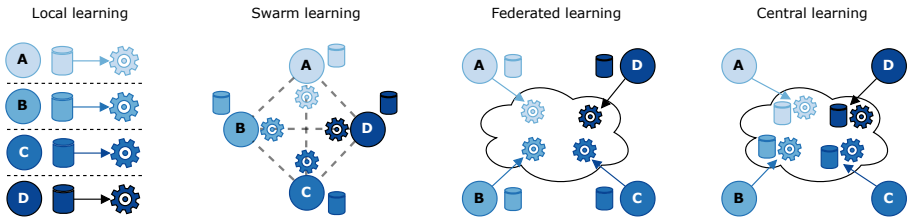


Figure 9.1: Visualisation of various machine learning architectures (based on (Warnat-Herresthal et al., 2021)). In the local learning scenario both data and parameters remain at the client. Federated learning only shares model parameters, whereas swarm learning removes the need for a centrally coordinating server and shares model parameters over blockchain while keeping data at client nodes. For central learning, both data and parameters are shared with a centrally coordinating server.

Decentralised machine learning could be the key towards privacy-preserving, trustworthy educational analytics (Ekuban and Domingue, 2023). Yet, only a couple of studies have investigated this promising area. Guo and Zeng (2020) use federated learning in the context of educational data analysis. They consider the task of dropout prediction in the KDD Cup 2015 dataset, achieving accuracy within a couple of percentage points of the central learning scenario. However, the authors do not make their code available and do not report performance metrics other than a single figure showing accuracy progression over epochs. This concern about their work was voiced by a more recent fed-

erated learning paper using the KDD Cup 2015 dataset. Fachola et al. (2023) achieve an accuracy of 81.7% in the case of central learning and show that using federated learning an accuracy of around 80% can be achieved, even when data is spread over more than 50 clients. A downside is that the reported accuracy of 81.7% is only two percentage points higher than the proportion of dropouts in the dataset of 79.3%. Accuracy is not the right choice of metric for this dataset. If we want to draw meaningful conclusions about the potential of federated learning analytics, we need to consider multiple datasets and performance metrics.

9.3 METHODOLOGY

This section describes the metrics we used to compare the performance of different models, the three datasets (OULAD, EdNet, and KDD Cup 2015) employed in our experiments, and the details of our federated learning algorithm.

9.3.1 *Metrics*

Two commonly used metrics to evaluate model performance are accuracy and F_1 score. Accuracy represents the fraction of correctly predicted records. The F_1 score is the harmonic mean of precision p (true positives divided by all predicted positives) and recall r (true positives divided by all actual positives). Both metrics should be used with caution when dealing with imbalanced datasets, as they are influenced heavily by whether the majority class is appointed as the positive or negative class.

A metric that is less explicitly sensitive to class imbalance is the Area Under the ROC Curve (AUC). The curve in question is a plot of the true positive rate (equal to recall) on the y-axis and the false positive rate (false positives divided by all actual negatives) on the x-axis. The curve is drawn by determining the true positive rate and the false positive rate at different classification thresholds, meaning AUC requires the probability estimates of a model for its calculation. Because AUC is based on probability outputs, rather than the 0-1 classification output, it can provide more fine-grained insight into whether a model is truly learning to separate positive from negative instances. AUC does suffer from its own issues, such as that it can be biased towards certain classifiers.

9.3.2 *Datasets*

The Open University Learning Analytics Dataset (OULAD) (Kuzilek et al., 2017), contains demographic data on students and logs of student activity within a virtual learning environment. The outcome variable of interest is

Table 9.1: Descriptive statistics of the three datasets we investigate in this chapter: OULAD, EdNet, and KDD Cup 2015. We additionally indicate state-of-the-art (SOTA) results for each, where the OULAD metrics are divided into PF (pass-fail), PW (pass-withdrawn), FD (fail-distinction), and PD (pass-distinction).

	OULAD (Waheed et al., 2020)	EDNET (D. Shin et al., 2021)	KDD CUP 2015 (W. Feng et al., 2019)
Use case	learning outcome	question correctness	dropout
# Students	32,593	784,309	200,902
# Records	10,655,280	95,293,926	13,545,124
% Pos. class	PF: 31% fail PW: 40% withdrawn FD: 30% distinction PD: 20% distinction	66% correct	79% dropout
SOTA	PF: Acc.=0.845 F_1=0.719 PW: Acc.=0.947 F_1=0.943 FD: Acc.=0.864 F_1=0.770 PD: Acc.=0.805 F_1=0.749	Acc.=0.725 AUC=0.791	F_1=0.929 AUC=0.909

the result a student achieved for a course, which can be pass, distinction, fail, or withdrawal. OULAD forms the basis for studies varying from the creation of predictive models identifying at-risk students (Hlosta et al., 2017) to the investigation of the role of demographics in virtual learning environments (Rizvi et al., 2019). We use the work of Waheed et al. (Waheed et al., 2020) as our baseline for comparison, as the authors provide a detailed description of the features they use, allowing us to conduct a replication that closely matches their process. They turn the original classification problem with four potential outcomes into four separate binary classification tasks (pass=0 & fail=1, pass=0 & withdrawn=1, fail=0 & distinction=1, pass=0 & distinction=1). Table 9.1 reports the accuracy and F_1 score achieved for each of these tasks.

EdNet is a knowledge tracing dataset containing data from users of a self-study platform (Choi, Y. Lee, D. Shin, et al., 2020). Rather than having a single outcome variable per user, EdNet involves predicting for each completed multiple-choice question whether a user answered it correctly. The prediction task of EdNet is temporal in nature, explaining why papers tackling this dataset tend to employ time-series machine learning models such as transformers (Choi, Y. Lee, J. Cho, et al., 2020). We use the SAINT+ transformer model (D. Shin et al., 2021) as our baseline for comparison, as this is the model with the current state-of-the-art performance. The authors use a version of EdNet with newer user data that is not publicly available. Yet, since the prediction task and features are identical, their results can still serve as a useful benchmark.

The final dataset we consider was used for the KDD Cup 2015 challenge. This dataset contains information on student interactions within a Massive Open Online Course (MOOC) environment. The goal is to predict student dropout, with a distinguishing characteristic being that 79% of the enrolled students dropped out. The dataset is thus highly imbalanced, explaining why KDD Cup 2015 papers tend to focus on reporting AUC and F_1 scores, rather than accuracy (W. Feng et al., 2019; W. Li et al., 2016).

9.3.3 *Federated learning*

Federated learning was proposed as a communication-efficient way to use all available data on individual devices to train a global model, without users having to share their personal data (McMahan et al., 2017). The use case considered when introducing swarm learning was that of a group of hospitals working together to create better predictive models for the detection of illnesses (Warnat-Herresthal et al., 2021) . The sensitivity of health data, along with the extensive legislation limiting data sharing in medical settings, provides a clear motivation for the need for a parameter-sharing infrastructure without a centrally coordinating party. A recent study in the educational field investigated a transfer learning approach and voiced concerns regarding the relevance of decentralised approaches for education (Gardner et al., 2023). Hence, we should ask to what extent decentralised machine learning contexts appear in educational environments.

Guo and Zeng (2020) and Fachola et al. (2023) envision a network of schools that are part of a federation sharing model parameters. These schools are part of the same governing body, but have separate physical locations, possibly even in different countries. From a legal and privacy perspective, it can then be worthwhile to employ federated learning to obtain optimal insight into student behaviour without needing to share student data across schools. The use case considered in both papers is dropout prediction using the KDD Cup 2015 dataset, meaning each student has a single outcome variable per course. Federated learning on the level of the classroom or the individual is likely not realistic here, since the majority of students have fewer than five course outcomes to train on. For the KDD Cup 2015 dataset we will therefore investigate federated learning performance up to a maximum of 100 local clients, corresponding to roughly 2,000 students per client. OULAD is comparable to the KDD Cup 2015 dataset, with the exception that it additionally contains demographic information. For OULAD we similarly analyse up to 100 local clients, corresponding to roughly 300 students per client.

For the EdNet setting, where a single student can answer thousands of questions in their self-study process, federated learning with individual students as local clients is more realistic. Nevertheless, since single users potentially have only one answered question within EdNet, it is not algorithmically practical to have local clients comprising one user. In our experiments, we

Table 9.2: Comparison of our central learning results to the results of Table 9.1, where the value between brackets represents the performance difference with earlier work.

	OULAD		EdNet		KDD Cup 2015	
	ACC.	F_1	ACC.	AUC	F_1	AUC
PF	0.862 (+0.017)	0.751 (+0.032)	0.720 (-0.005)	0.757 (-0.035)	0.925 (-0.003)	0.881 (-0.028)
PW	0.933 (-0.014)	0.914 (-0.011)				
FD	0.893 (+0.029)	0.820 (+0.050)				
PD	0.810 (+0.005)	0.199 (-0.551)				

will investigate the performance of local and federated learning up to a maximum of 100 local clients, corresponding to around 100 users per client when working with a randomly selected subset of 10,000 students.

9.4 RESULTS

The Python code used to produce the outcomes of this section and detailed results per dataset are available on GitHub[1]. Our federated learning code adheres to the `FedAvg` algorithm of McMahan et al. (2017). Central learning experiments were conducted using the machine learning library scikit-learn and the gradient boosting libraries XGBoost and CatBoost. We used Pytorch as the deep learning library for our federated learning algorithm and exclusively used XGBoost with default settings as our local learning classifier.

9.4.1 *Central learning*

Table 9.2 presents our central learning results using 10-fold cross-validation with an 80-20 train-test split. Our best results were achieved using CatBoost (OULAD and KDD Cup 2015) and XGBoost (EdNet). Table 9.2 shows that we managed to achieve comparable performance to the current state-of-the-art.

Since Waheed et al. (2020) extensively describe the features they engineered, we were able to reproduce these features and use them as input for OULAD classification. For the EdNet prediction task, we created lag features for previous user question correctness to turn the time series prediction task into a classification task. This enabled us to utilise the regular machine learning and gradient boosting libraries we used for OULAD and KDD Cup 2015. For the KDD Cup 2015 dataset, we designed student activity features similar to those of OULAD.

1 https://github.com/MaxvanHaastrecht/Federated-Learning-Analytics

9.4.2 *Local learning and federated learning*

For our local and federated learning scenarios, we divided students randomly over clients. For OULAD federated learning, we used a neural network with two hidden layers of sizes 30 and 10, a learning rate η of 0.02, a cross-entropy loss function with the Adam optimiser, the number of communication rounds R set to 50, the number of local epochs per round $E = 2$, and a batch size of 64. Figure 9.2 shows that both federated learning and local learning perform worse than the central learning scenario. However, whereas local learning accuracy drops significantly as we progress from 10 to 100 local clients, federated learning accuracy remains roughly constant.

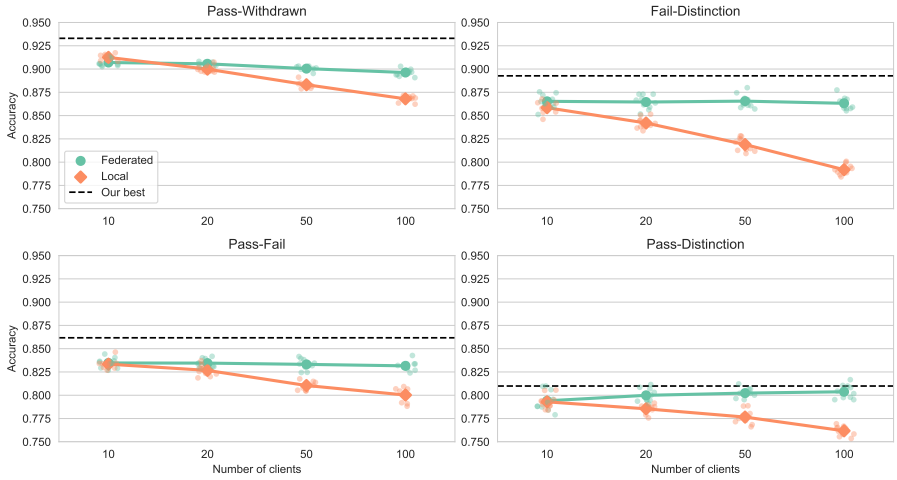


Figure 9.2: Plot of the bootstrapped mean accuracy for varying numbers of local clients, showing comparisons of our local learning, federated learning, and central learning results.

Figure 9.3 summarises the results from our EdNet and KDD Cup 2015 experiments. For KDD Cup 2015, we used the exact same federated learning settings as with OULAD. For EdNet, we changed the batch size to 128, as is used in earlier work (Choi, Y. Lee, J. Cho, et al., 2020), and lowered the number of communication rounds R from 50 to 20. We additionally used hidden layer sizes of 16 and 8, rather than 30 and 10, since EdNet feature engineering resulted in fewer input features for the network. Since the EdNet dataset is comparatively large, it is common practice to work with a random subset of the dataset in experimental settings such as our federated learning context (Long et al., 2022; Y. Yang et al., 2021). We work with a random subset of 10,000 students and indicate the AUC of our best central learning model in Figure 9.3.

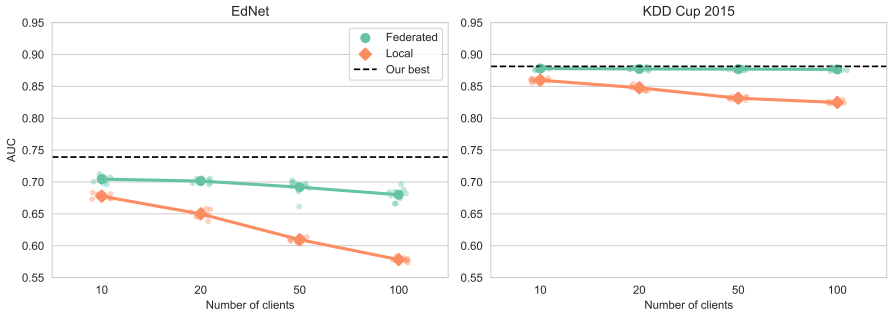


Figure 9.3: Plot of the bootstrapped mean AUC for varying numbers of local clients, showing comparisons of our central learning EdNet and KDD Cup 2015 AUC results to local learning and federated learning.

9.4.3 *Federated learning analytics metric (FLAME)*

Our numerical results provide an indication of the performance of federated learning compared to local learning and central learning. However, our results are not directly usable by policymakers in education deciding whether to opt for a federated learning architecture. Questions remain regarding the optimal number of local clients in each scenario and how much performance we are willing to trade off for an improved preservation of privacy. To ease the decision-making process, we propose the federated learning analytics metric (FLAME). The idea behind FLAME is to capture the trade-off between privacy and performance in a single metric, such that comparisons across scenarios, datasets, and numbers of local clients become more tenable. We define FLAME as:

$$\text{FLAME} = \frac{1 - \frac{1}{K}}{1 + (p_c - p_f)} = \frac{\text{privacy gain}}{1 + \text{performance loss}},$$

where K is the number of local clients, p_c is the central learning performance, and p_f is the federated learning performance. For institutions considering to move from a central learning architecture to federated learning, p_c will be a known quantity. For institutions that do not have a centralised architecture, p_c can be estimated based on the literature or through simulations. FLAME is suited to be used for performance metrics ranging between [0,1], such as accuracy, F_1, and AUC. The numerator captures the gain in privacy achieved by employing an architecture with local clients. The denominator captures the loss in performance.

Figure 9.4 shows the FLAME values for EdNet and KDD Cup 2015, where AUC is the relevant performance metric. FLAME values for the local learning scenario are also shown, which can be calculated by replacing the federated learning performance in the FLAME formula with local learning performance.

Taking EdNet as an example, we observe that for federated learning FLAME peaks at 50 clients, whereas for local learning FLAME peaks at 20 clients. By more explicitly incorporating the privacy-performance trade-off, FLAME therefore clarifies differences between algorithms in a way the pure AUC scores of Figure 9.3 cannot.

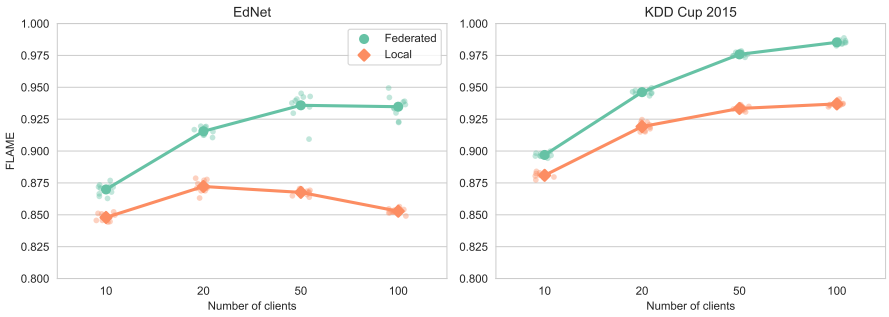


Figure 9.4: FLAME values for EdNet and KDD Cup 2015, where AUC is the performance metric. In the case of 50 local clients, AUC loss must be less than 0.0315 to achieve a FLAME higher than 0.95.

9.5 DISCUSSION

Our results demonstrate the potential of federated learning to preserve privacy and performance in educational contexts. For OULAD, we observed that our federated learning algorithm achieved comparable accuracy to earlier results for three out of four scenarios considered, even when the number of local clients was set to 100. For the KDD Cup 2015 dataset, federated learning matched our best results, again up to 100 local clients. Federated learning also significantly outperformed local learning for all three datasets. When dividing data over 100 local clients, the average accuracy gain for OULAD was 4.32% and the average AUC gains for EdNet and KDD Cup 2015 were 0.1017 and 0.0518, respectively.

Our FLAME values in Figure 9.4 demonstrated that local learning and federated learning warrant serious consideration in settings where dividing data over 20 or more clients is realistic. However, the answer to student privacy concerns can never be purely technological. Federated learning is promising, but it carries with it additional security risks and questions whether student's perceptions of these technologies are as positive as their theoretical benefits. Yet, given the increasing tensions between the datafication of education and the privacy concerns of students, privacy-preserving machine learning architectures may offer the path of least resistance towards a bright future for educational analytics.

Federated learning is perhaps the most commonly used privacy-preserving machine learning strategy, but certainly not the only one. We did not cover other paradigms within this chapter, such as split learning (Thapa et al., 2022), swarm learning (Warnat-Herresthal et al., 2021), and transfer learning (Gardner et al., 2023). In future work, it will be crucial to compare the privacy-performance trade-off for various approaches. We should be aware that in contexts where performance takes precedent, combining strategies (e.g., federated learning and split learning (Thapa et al., 2022)) might be the optimal choice, whereas in contexts where privacy is paramount, a local learning approach that fosters stakeholder trust could provide the perfect fit. Regardless of the privacy-preserving paradigms considered, insights regarding the privacy-performance trade-off provided by FLAME can serve as a useful starting point for discussion.

A limitation of our work is that all benchmarking datasets had drawbacks. OULAD is extensively documented and publicly available, but is comprised of scenarios with imbalanced classification tasks where the metrics currently used in the literature (accuracy and F_1) are inadequate for thorough comparisons of model performance. EdNet is publicly available, but recent work has relied on a version of the dataset that is not publicly available (D. Shin et al., 2021), or has worked with subsets of the full dataset that hinder replicability (Long et al., 2022; Y. Yang et al., 2021). The KDD Cup 2015 dataset is not publicly available from a dedicated website, and the most relevant publications covering this dataset in recent years only report model accuracy (Fachola et al., 2023; Guo and Zeng, 2020), when this is a highly imbalanced dataset with 79% of students dropping out. These drawbacks are not ideal, but we strongly believe these datasets offer an accurate representation of currently available benchmarks. Still, we require better benchmark datasets and accompanying research in the future.

9.5.1 *Interviews with stakeholders*

To uncover the views of stakeholders at educational institutions regarding federated learning, we plan to conduct a follow-up study where we use a grounded theory approach to analyse the data resulting from a series of qualitative interviews. The analysis of the first two interviews with educational technology experts in higher education have been completed at this stage, and we deem it relevant to report two preliminary findings here.

Firstly, the experts we interviewed pointed out that federated learning could serve as a stepping stone for educational institutions to move from experimental situations to wide-scale impact. Interestingly, the two experts both used the metaphor of a chicken and egg situation, whereby a prerequisite to scale up an educational innovation is a demonstration of its impact, but to demonstrate impact you need the data of students that you only get after you scale up. One of the interviewees put it as follows: "It's kind of chicken

and egg situation. To demonstrate that an algorithm can be trusted, you have to do some kind of analysis, so people can see that it offers advantages and makes education better. But then you have to be able to start, and if there is suspicion regarding an innovation then you can never start anything." Federated learning could help to perform the required analysis without immediately having to implement a solution that is not trusted by students.

Secondly, one of the interviewees explained why they consider it worthwhile to keep developing privacy-preserving machine learning techniques with regards to the concept of proportionality: "If you have no other option than central learning, then you can talk all you want about proportionality, but then you have no choice. If you can use different methods, you can try to find a balance in privacy risk and usability." In other words, if we do not keep developing privacy-preserving machine learning techniques for education, cases will occur where our only realistic option is central learning. We will then find ourselves in a situation where we cannot adequately attend to the proportionality principle which is central to regulations such as GDPR.

9.6 CONCLUSION AND FUTURE WORK

With education becoming more datafied than ever, researchers interested in optimising learning environments are increasingly faced with questions regarding ethics, privacy, fairness, and trustworthiness. Decisions to intrude on student privacy should be taken with the utmost caution. There are legitimate concerns whether any type of optimisation warrants the collection of sensitive learner data. Within this context, privacy-preserving machine learning that respects privacy while maintaining model performance is an intriguing recent development. However, until now, we lacked rigorous investigations of the impact of privacy-preserving architectures on educational analytics model performance.

We compared algorithm performance across three architectures (local learning, federated learning, central learning) for three different prediction use cases (learning outcome, question correctness, dropout). In doing so, we provided a comprehensive image of what can be achieved with privacy-preserving architectures. We found that even when dividing data over 100 clients, federated learning can compete with state-of-the-art results. A major finding was that although for 50 or more clients federated learning outperformed local learning, differences were often not significant when dividing data over 20 or fewer clients. This points to the importance of considering local learning as a privacy-preserving strategy for educational analytics. Future work will need to extend the investigation of how students, teachers, and other stakeholders view federated learning, since the relative complexity of privacy-preserving machine learning may diminish trust. Nevertheless, as evidenced by the preliminary findings from our interviews with stakeholders,

the datafication of education combined with the clear wish of students to preserve privacy signal a promising future for federated learning analytics.

10

CONCLUSION: TRANSDISCIPLINARY PERSPECTIVES ON VALIDITY

At the outset of this dissertation, we motivated why a transdisciplinary research approach could potentially benefit the design and validation of technology-enhanced learning (TEL) solutions, highlighting its specific relevance in connection to our GEIGER cybersecurity project for SMEs. We presented our main research question: *How can transdisciplinary research inform the design and validation of technology-enhanced learning solutions?* In this concluding chapter, we will reflect on how the individual pieces of our research puzzle have helped us move towards an answer to our main research question. Additionally, we will discuss how our designed artefacts have impacted science and society, and will contemplate possibly fruitful directions for future research.

10.1 CONTRIBUTIONS

Figure 10.1, first presented in Chapter 1, visualises the research process that was followed in this dissertation. We matched the phases of the transdisciplinary research process, which informed how we could incorporate insights from different fields of research and societal stakeholders, to the phases of the engineering cycle, which informed the research methods we used to answer the research questions of the various chapters in this dissertation. We can now reflect on how the answers to sub-questions combine towards answering our main research question.

CHAPTER 2 uncovered the elements of an accessible and swift systematic review methodology. We presented the systematic review methodology SYMBALS, which combines an active learning approach in the title and abstract screening phase with a backward snowballing step to find additional literature. Using two case studies, we demonstrated the ability of SYMBALS to speed up the review process, while simultaneously managing to retrieve a significant proportion of all relevant papers. SYMBALS was used in several later chapters within this dissertation, and has been used by scientists to aid their systematic review process in fields ranging from computer science to marine policy to sports medicine. Thus, this chapter formed the first step in investigating our problem domain.

CHAPTER 3 examined the topic of SME cybersecurity measurement using a SYMBALS systematic review. We synthesised our findings into a socio-technical cybersecurity framework for SMEs, where we indicated how different cornerstones of the SME socio-technical system can be expected to interact at different levels of digital maturity. The framework developed in Chapter 3 informed the co-creation and design work we executed in later studies.

CHAPTER 4 addressed the question: How should an SME cybersecurity application be designed to motivate users? Using a collaborative design research approach, we designed an initial version of the GEIGER application. The educational content and user interface of the application were created together with users, and were also informed by the cybersecurity framework of Chapter 3 and behavioural theories. The artefact resulting from this study, a prototype educational cybersecurity application for SMEs, is central to the GEIGER product offering to this day. Through the active involvement of SME users, this study represented the initial foray into the domain of transdisciplinary research. Regarding our main research question, we can surmise from the findings of this chapter that a transdisciplinary approach to the design of TEL solutions can help to promote motivation by explicitly considering the behavioural needs of users.

CHAPTER 5 used a technical action research approach to investigate how cyber threat intelligence could be incorporated into the GEIGER application. We described in detail how threat intelligence, which in its raw form can be difficult to understand for cybersecurity experts, could be turned into actionable insights for SMEs. The threat intelligence platform that we developed together with industry partners was the first example of a technical cybersecurity pipeline that provided real-time, understandable insights to users with limited cybersecurity knowledge and resources. Through the involvement of both industry partners and SMEs, we learned that just because raw data is considered too difficult to understand for people without expert knowledge, it does not mean this raw data cannot be used to create an improved design for these people.

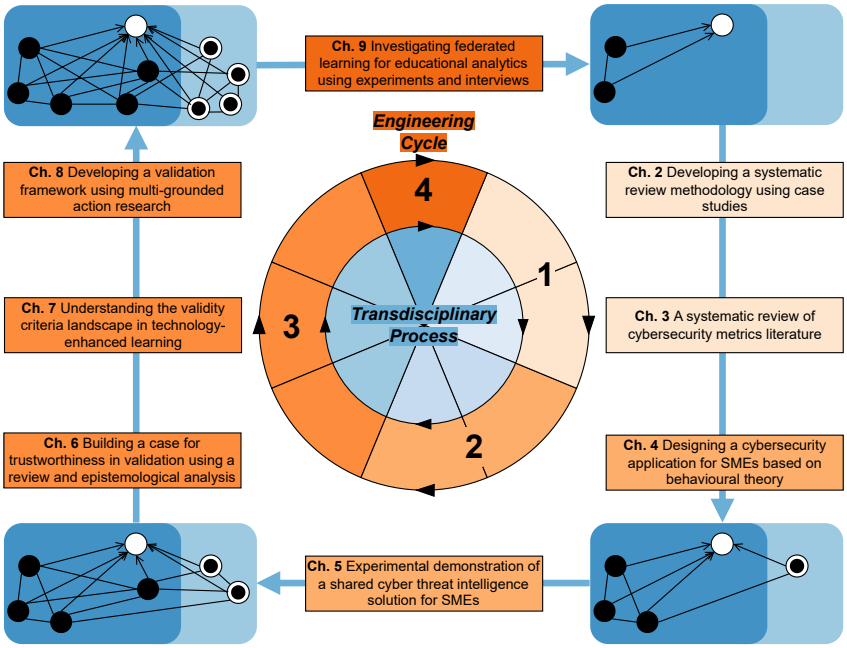


Figure 10.1: The visualisation of our research process that was first presented in Chapter 1. We combine the transdisciplinary process described by Lawrence et al. (2022) and the engineering cycle of Wieringa (2014).

CHAPTER 6 was the first chapter where our focus shifted from design to validation, and from a narrow, context-specific view used to design an educational cybersecurity application for SMEs, to a broad view used to develop a validation framework for TEL. In Chapter 6, we employed a combination of a literature review and an epistemological analysis to develop a theoretical basis

for validity considerations in learning analytics. We presented an overview of how existing validity criteria are used by researchers, which informed the design of a Learning Analytics Validation Assistant (LAVA).

CHAPTER 7 extended the work of Chapter 6 using a SYMBALS systematic review. We uncovered which validity criteria are considered in TEL research, which methods are used to gain insight into these criteria, and whether criteria are on average assessed positively or negatively. By comparing criteria definitions and usage over time, we created an overview of the validity criteria landscape, which could inform future holistic validation frameworks. In combination, Chapter 6 and Chapter 7 demonstrated how crossing disciplinary boundaries can yield a more holistic image of validity in the context of TEL.

CHAPTER 8 investigated how a holistic validation framework for TEL could be constructed. Through a multi-grounded action research approach we developed VAST, a validation framework for e-assessment technologies such as GEIGER. We additionally created a guideline to accompany our academic contribution (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023), which is intended to help users of VAST gain an understanding of the step-by-step process underlying the framework. Whether our framework will serve as a useful validation tool for researchers and practitioners is yet to be seen, but the societal stakeholders with which we developed the framework have surely gained valuable insights concerning potential validation strategies. The input from societal stakeholders helped us to gain an understanding of the importance of clarity and flexibility in validation frameworks; understanding we would not have been able to gain without a transdisciplinary research approach.

CHAPTER 9 covered the question: How does the privacy-performance trade-off manifest itself in educational analytics? After performing technical experiments to demonstrate the potential of federated learning for educational analytics, we introduced a novel metric (FLAME) that assists policymakers in their assessment of the privacy-performance trade-off. We presented preliminary findings from a series of interviews with stakeholders, to reflect on the viability of introducing advanced machine learning techniques into educational contexts. The interviewees indicated that federated learning could serve as a stepping stone to move from experimental techniques to large-scale innovation, whereas we had initially envisioned a federated learning architecture as a replacement for central learning. This formed another reminder that a transdisciplinary research approach can not only inform the comprehensive validation of TEL innovations, but might in fact be a requirement for comprehensive validation.

10.2 IMPLICATIONS

Our transdisciplinary research approach informed the design and validation of the GEIGER solution. But can we generalise our findings beyond the GEIGER project?

We demonstrated how technical knowledge extracted from scientific literature using an innovative systematic review approach (Chapter 2, Chapter 3), can be incorporated in the design of a TEL solution in collaboration with users and industry partners (Chapter 4, Chapter 5). We argue that although the work of the first chapters focused primarily on the GEIGER use case, its findings are applicable to a large range of contexts, as exemplified by the variety of research areas in which SYMBALS has been employed.

Later chapters exercised a broader view from the outset, to inform the studies related to validation. We developed a theoretical basis for learning analytics validation (Chapter 6), before expanding on this work using a SYMBALS review to create a comprehensive overview of the TEL validity landscape (Chapter 7). We designed a comprehensive framework for e-assessment technologies (Chapter 8), and applied our theoretical validation knowledge in an evaluation study of privacy-preserving machine learning for educational analytics (Chapter 9). Although the work of these later chapters was predominantly theoretical, the accumulated knowledge was generally developed in collaboration with the societal partners of the GEIGER project. We believe that our transdisciplinary approach increased the potential of our validity theory contributions to create an impact in the wider TEL domain.

Focusing on our main research question, we can conclude that transdisciplinary research facilitates the discovery of practical barriers to successfully implementing existing TEL methods, frameworks, and artefacts. However, transdisciplinary research also opens our eyes to how we can adapt and enhance our current solutions to better cater to the needs of society. Whether it is through more adequately addressing the behavioural and pedagogic needs of users, or through more critically reflecting on and contextualising our validity evidence, building bridges between science and society introduces us to new perspectives that positively influence the design and validation of TEL solutions.

10.3 LIMITATIONS

The chapters of this dissertation each mention the limitations of their corresponding studies. Three further overarching limitations should be mentioned here. Firstly, the nature of GEIGER as a research and innovation project had as a consequence that its process was fast-paced. Practical progress was regularly swifter than that of the accompanying scientific research. The result was that although the connection between science and society was prominent within the GEIGER project, it was not always as prominent within the scientific stud-

ies of this dissertation. The process of collaboration with users and industry partners is primarily described in project deliverables, which may limit the clarity regarding the impact of our transdisciplinary approach within this dissertation.

Secondly, the GEIGER project took place during the COVID pandemic. The project was luckily able to move forward, but with the limitation that much fewer personal interactions with users and project partners took place than we had planned for. We have highlighted the importance of thick descriptions of educational contexts that allow for the critical contextualisation of validity evidence. The COVID pandemic limited our ability to critically contextualise. However, we continuously sought contact with users and partners online, and used the few opportunities for in-person interaction as effectively as we could, while nonetheless remaining aware of the impact the pandemic had on our research.

Finally, one can ask to what extent we managed to achieve a satisfactory answer to our main research question. In one sense, we can argue that we have uncovered several ways in which transdisciplinary research can inform the design and validation of TEL solutions, and have thus provided an answer to our main research question. However, in another sense, certain questions remain open and we cannot rule out the possibility that there are ways in which transdisciplinary research can positively impact TEL design and validation beyond those presented here. This can be interpreted as a limitation of this dissertation, but can also be understood as a gap for future research to address.

10.4 FUTURE DIRECTIONS

Because of the nature of the engineering cycle and the transdisciplinary research process, suggestions for future directions, for a future cycle, come primarily from the studies positioned towards the end of the current cycle. Many of the questions posed in the final chapters remain open. We need to continue to adapt validation frameworks to novel technological developments, such as those producing process data in educational environments. We have to clarify existing validation frameworks and increase their flexibility, such that they become more usable for researchers and practitioners. Additionally, we should continue to unearth diverse stakeholder perspectives on our designed solutions, if we are to legitimately recognise the diversity of perspectives that exist regarding technological innovations in education. We can surmise that TEL validity theory offers a promising direction for future research endeavours.

To close, we want to reflect on two of the findings from Chapter 7, and the necessity to further investigate their implications. We observed a correlation between the research method used to assess validity criteria and the outcome - negative, positive, or mixed - of that assessment. We also exposed a potentially

problematic hierarchy in validity criteria, where certain criteria receive a much higher priority than others. If our validation strategies are misguided, our innovations will follow this misguided path. We cannot accept such a future, and thus we will need to investigate where our validation strategies may be heading astray, such that we can correct our course. Albert Einstein once said: "Not everything that can be counted counts, and not everything that counts can be counted." Let us, as science and society, figure out what counts.

BIBLIOGRAPHY

Abdulrahaman, M. D., Faruk, N., Oloyede, A. A., Surajudeen-Bakinde, N. T., Olawoyin, L. A., Mejabi, O. V., Imam-Fulani, Y. O., Fahm, A. O., and Azeez, A. L. (2020). "Multimedia Tools in the Teaching and Learning Processes: A Systematic Review." In: *Heliyon* 6.11. DOI: 10.1016/j.heliyon.2020.e05312.

Abe, S., Uchida, Y., Hori, M., Hiraoka, Y., and Horata, S. (2018). "Cyber Threat Information Sharing System for Industrial Control System (ICS)." In: *2018 57th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. Nara, Japan: IEEE, pp. 374–379. DOI: 10.23919/SICE.2018.8492570.

Addey, C., Maddox, B., and Zumbo, B. D. (2020). "Assembled Validity: Rethinking Kane's Argument-Based Approach in the Context of International Large-Scale Assessments (ILSAs)." In: *Assessment in Education: Principles, Policy & Practice* 27.6, pp. 588–606. DOI: 10.1080/0969594X.2020.1843136.

Ahadi, A., Bower, M., Lai, J., Singh, A., and Garrett, M. (2021). "Evaluation of Teacher Professional Learning Workshops on the Use of Technology - a Systematic Review." In: *Professional Development in Education* 50.1, pp. 221–237. DOI: 10.1080/19415257.2021.2011773.

Albakri, A., Boiten, E., and De Lemos, R. (2018). "Risks of Sharing Cyber Incident Information." In: *Proceedings of the 13th International Conference on Availability, Reliability and Security*. ARES 2018. Hamburg, Germany: Association for Computing Machinery, pp. 1–10. DOI: 10.1145/3230833.3233284.

Alberts, C. J., Dorofee, A. J., Stevens, J. F., and Woody, C. (2005). *OCTAVE-S Implementation Guide, Version 1*. Tech. rep. Software Engineering Institute, Carnegie Mellon University.

Alencar Rigon, E., Merkle Westphall, C., Ricardo dos Santos, D., and Becker Westphall, C. (2014). "A Cyclical Evaluation Model of Information Security Maturity." In: *Information Management & Computer Security* 22.3, pp. 265–278. DOI: 10.1108/IMCS-04-2013-0025.

AlHogail, A. (2015). "Design and Validation of Information Security Culture Framework." In: *Computers in Human Behavior* 49, pp. 567–575. DOI: 10.1016/j.chb.2015.03.054.

Ali, L., Hatala, M., Gašević, D., and Jovanović, J. (2012). "A Qualitative Evaluation of Evolution of a Learning Analytics Tool." In: *Computers & Education* 58.1, pp. 470–489. DOI: 10.1016/j.compedu.2011.08.030.

Allodi, L. and Massacci, F. (2017). "Security Events and Vulnerability Data for Cybersecurity Risk Estimation." In: *Risk Analysis* 37.8, pp. 1606–1627. DOI: 10.1111/risa.12864.

Aloisi, C. (2023). "The Future of Standardised Assessment: Validity and Trust in Algorithms for Assessment and Scoring." In: *European Journal of Education* 58.1, pp. 98–110. DOI: 10.1111/ejed.12542.

Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., and Fernández-Manjón, B. (2019). "Lessons Learned Applying Learning Analytics to Assess Serious Games." In: *Computers in Human Behavior* 99, pp. 301–309.

Alpcan, T. and Bambos, N. (2009). "Modeling Dependencies in Security Risk Management." In: *2009 Fourth International Conference on Risks and Security of Internet and Systems (CRiSIS 2009)*, pp. 113–116. DOI: 10.1109/CRISIS.2009.5411969.

Ansari, M. S., Bartos, V., and Lee, B. (2020). "Shallow and Deep Learning Approaches for Network Intrusion Alert Prediction." In: *Procedia Computer Science*. Third International Conference on Computing and Network Communications (CoCoNet'19) 171, pp. 644–653. DOI: 10.1016/j.procs.2020.04.070.

Asghar, M. R., Hu, Q., and Zeadally, S. (2019). "Cybersecurity in Industrial Control Systems: Issues, Technologies, and Challenges." In: *Computer Networks* 165, p. 106946. DOI: 10.1016/j.comnet.2019.106946.

Atamli, A. W. and Martin, A. (2014). "Threat-Based Security Analysis for the Internet of Things." In: *2014 International Workshop on Secure Internet of Things*, pp. 35–43. DOI: 10.1109/SIoT.2014.10.

Azad, M. A., Bag, S., Ahmad, F., and Hao, F. (2021). "Sharing Is Caring: A Collaborative Framework for Sharing Security Alerts." In: *Computer Communications* 165, pp. 75–84. DOI: 10.1016/j.comcom.2020.09.013.

Baars, T., Mijnhardt, F., Vlaanderen, K., and Spruit, M. (2016). "An Analytics Approach to Adaptive Maturity Models Using Organizational Characteristics." In: *Decision Analytics* 3.1, p. 5. DOI: 10.1186/s40165-016-0022-1.

Babineau, J. (2014). "Product Review: Covidence (Systematic Review Software)." In: *Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada* 35.2, pp. 68–71. DOI: 10.5596/c14-016.

Badri, S., Fergus, P., and Hurst, W. (2016). "Critical Infrastructure Automated Immuno-Response System (CIAIRS)." In: *2016 International Conference on Control, Decision and Information Technologies (CoDIT)*. Saint Julian's, Malta: IEEE, pp. 096–101. DOI: 10.1109/CoDIT.2016.7593542.

Badsha, S., Vakilinia, I., and Sengupta, S. (2019). "Privacy Preserving Cyber Threat Information Sharing and Learning for Cyber Defense." In: *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*. Las Vegas, NV, USA: IEEE, pp. 0708–0714. DOI: 10.1109/CCWC.2019.8666477.

Baesso Moreira, G., Menditi Calegario, V., Duarte, J. C., and F. Pereira dos Santos, A. (2018). "Extending the VERIS Framework to an Incident Handling Ontology." In: *2018 IEEE/WIC/ACM International Conference on Web*

Intelligence (WI). Santiago, Chile: IEEE, pp. 440–445. DOI: 10.1109/WI.2018.00-55.

Barnum, S. (2012). *Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression (STIX)*. Technical Paper. Mitre Corporation.

Barrett, M. P. (2018). *Framework for Improving Critical Infrastructure Cybersecurity Version 1.1*. Tech. rep. NIST.

Başağaoğlu Demirekin, Z. and Buyukcavus, M. H. (2022). "Effect of Distance Learning on the Quality of Life, Anxiety and Stress Levels of Dental Students during the COVID-19 Pandemic." In: *BMC Medical Education* 22.1, pp. 1–9. DOI: 10.1186/s12909-022-03382-y.

Baskerville, R. and Pries-Heje, J. (1999). "Grounded Action Research: A Method for Understanding IT in Practice." In: *Accounting, Management and Information Technologies* 9.1, pp. 1–23. DOI: 10.1016/S0959-8022(98)00017-4.

Bassett, G., Hylender, C. D., Langlois, P., Pinto, A., and Widup, S. (2021). *2021 Data Breach Investigations Report*. Technical Report. Verizon.

Ben Mahmoud, M. S., Larrieu, N., and Pirovano, A. (2011). "A Risk Propagation Based Quantitative Assessment Methodology for Network Security - Aeronautical Network Case Study." In: *2011 Conference on Network and Information Systems Security*, pp. 1–9. DOI: 10.1109/SAR-SSI.2011.5931372.

Bennett, R. E. and Bejar, I. I. (1998). "Validity and Automated Scoring: It's Not Only the Scoring." In: *Educational Measurement: Issues and Practice* 17.4, pp. 9–17. DOI: 10.1111/j.1745-3992.1998.tb00631.x.

Benz, M. and Chatterjee, D. (2020). "Calculated Risk? A Cybersecurity Evaluation Tool for SMEs." In: *Business Horizons* 63.4, pp. 531–540. DOI: 10.1016/j.bushor.2020.03.010.

Berman, N. B. and Artino, A. R. (2018). "Development and Initial Validation of an Online Engagement Metric Using Virtual Patients." In: *BMC Medical Education* 18.1, p. 213. DOI: 10.1186/s12909-018-1322-z.

Best, D. M., Bhatia, J., Peterson, E. S., and Breaux, T. D. (2017). "Improved Cyber Threat Indicator Sharing by Scoring Privacy Risk." In: *2017 IEEE International Symposium on Technologies for Homeland Security (HST)*. Waltham, MA, USA: IEEE, pp. 1–5. DOI: 10.1109/THS.2017.7943482.

Bhilare, D. S., Ramani, A., and Tanwani, S. (2008). "Information Security Assessment and Reporting: Distributed Defense." In: *undefined*.

Bissell, K. and Lasalle, R. M. (2019). *2019 Cost of Cybercrime Study*. Technical Report. Accenture and Ponemon Institute.

Bitner, R., Le, N.-T., and Pinkwart, N. (2020). "A Concurrent Validity Approach for EEG-Based Feature Classification Algorithms in Learning Analytics." In: *Proceedings of the 12th International Conference on Computational Collective Intelligence*. ICCCI'20. Da Nang, Vietnam: Springer-Verlag, pp. 568–580. DOI: 10.1007/978-3-030-63007-2_44.

Böhme, R. and Freiling, F. C. (2008). "On Metrics and Measurements." In: *Dependability Metrics: Advanced Lectures*. Ed. by I. Eusgeld, F. C. Freiling, and R. Reussner. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 7–13.

Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., and Kerres, M. (2020). "Mapping Research in Student Engagement and Educational Technology in Higher Education: A Systematic Evidence Map." In: *International Journal of Educational Technology in Higher Education* 17.1, pp. 1–30. DOI: 10.1186/s41239-019-0176-8.

Borah, R., Brown, A. W., Capers, P. L., and Kaiser, K. A. (2017). "Analysis of the Time and Workers Needed to Conduct Systematic Reviews of Medical Interventions Using Data from the PROSPERO Registry." In: *BMJ Open* 7.2, e012545. DOI: 10.1136/bmjopen-2016-012545.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., and Pereira, J. (2016). "An Update to the Systematic Literature Review of Empirical Evidence of the Impacts and Outcomes of Computer Games and Serious Games." In: *Computers & Education* 94, pp. 178–192. DOI: 10.1016/j.compedu.2015.11.003.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M. (2007). "Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain." In: *Journal of Systems and Software*. Software Performance 80.4, pp. 571–583. DOI: 10.1016/j.jss.2006.07.009.

Brewer, P. E., Racy, M., Hampton, M., Mushtaq, F., Tomlinson, J. E., and Ali, F. M. (2021). "A Three-Arm Single Blind Randomised Control Trial of Naïve Medical Students Performing a Shoulder Joint Clinical Examination." In: *BMC Medical Education* 21.1, pp. 1–7. DOI: 10.1186/s12909-021-02822-5.

Brinkhuis, M. J. S., Savi, A. O., Hofman, A. D., Coomans, F., Maas, H. L. J. van der, and Maris, G. (2018). "Learning As It Happens: A Decade of Analyzing and Shaping a Large-Scale Online Learning System." In: *Journal of Learning Analytics* 5.2, pp. 29–46. DOI: 10.18608/jla.2018.52.3.

Broniatowski, D. A. and Tucker, C. (2017). "Assessing Causal Claims about Complex Engineered Systems with Quantitative Data: Internal, External, and Construct Validity." In: *Systems Engineering* 20.6, pp. 483–496. DOI: 10.1002/sys.21414.

Brotsis, S., Kolokotronis, N., Limniotis, K., Shiaeles, S., Kavallieros, D., Bellini, E., and Pavué, C. (2019). "Blockchain Solutions for Forensic Evidence Preservation in IoT Environments." In: *2019 IEEE Conference on Network Softwarization (NetSoft)*. Paris, France: IEEE, pp. 110–114. DOI: 10.1109/NETSOFT.2019.8806675.

Brown, A. L. and Campione, J. C. (1996). "Psychological Theory and the Design of Innovative Learning Environments: On Procedures, Principles, and Systems." In: *Innovations in Learning: New Environments for Education*. 1st. Mahwah, NJ, US: Lawrence Erlbaum Associates, Inc, pp. 289–325.

Brown, S., Moye, T., Hubertse, R., and Glăvan, C. (2019). "Towards Mature Federated Cyber Incident Management and Information Sharing Capabilities in NATO and NATO Nations." In: *MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM)*. Norfolk, VA, USA: IEEE, pp. 1–5. DOI: 10.1109/MILCOM47813.2019.9020814.

Browning, K. (2021). "Up to 1,500 Businesses Could Be Affected by a Cyberattack Carried out by a Russian Group." In: *The New York Times*.

Brožová, H., Šup, L., Rydval, J., Sadok, M., and Bednar, P. (2016). "Information Security Management: ANP Based Approach for Risk Analysis and Decision Making." In: *AGRIS on-line Papers in Economics and Informatics* 08.1, pp. 1–11.

Buchanan, R. (1992). "Wicked Problems in Design Thinking." In: *Design Issues* 8.2, pp. 5–21. DOI: 10.2307/1511637. JSTOR: 1511637.

Burger, E. W., Goodman, M. D., Kampanakis, P., and Zhu, K. A. (2014). "Taxonomy Model for Cyber Threat Intelligence Information Exchange Technologies." In: *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security*. WISCS '14. Scottsdale, AZ, USA: Association for Computing Machinery, pp. 51–60. DOI: 10.1145/2663876.2663883.

Cadena, A., Gualoto, F., Fuertes, W., Tello-Oquendo, L., Andrade, R., Tapia, F., and Torres, J. (2020). "Metrics and Indicators of Information Security Incident Management: A Systematic Mapping Study." In: *Developments and Advances in Defense and Security*. Ed. by Á. Rocha and R. P. Pereira. Smart Innovation, Systems and Technologies. Singapore: Springer, pp. 507–519. DOI: 10.1007/978-981-13-9155-2_40.

Carías, J. F., Borges, M. R. S., Labaka, L., Arrizabalaga, S., and Hernantes, J. (2020). "Systematic Approach to Cyber Resilience Operationalization in SMEs." In: *IEEE Access* 8, pp. 174200–174221. DOI: 10.1109/ACCESS.2020.3026063.

Carías, J. F., Arrizabalaga, S., Labaka, L., and Hernantes, J. (2020). "Cyber Resilience Progression Model." In: *Applied Sciences* 10.21, p. 7393. DOI: 10.3390/app10217393.

Casola, V., De Benedictis, A., Rak, M., and Villano, U. (2019). "Toward the Automation of Threat Modeling and Risk Assessment in IoT Systems." In: *Internet of Things* 7, p. 100056. DOI: 10.1016/j.iot.2019.100056.

– (2020). "A Novel Security-by-Design Methodology: Modeling and Assessing Security by SLAs with a Quantitative Approach." In: *Journal of Systems and Software* 163, p. 110537. DOI: 10.1016/j.jss.2020.110537.

Cerro Martínez, J. P., Guitert Catasús, M., and Romeu Fontanillas, T. (2020). "Impact of Using Learning Analytics in Asynchronous Online Discussions in Higher Education." In: *International Journal of Educational Technology in Higher Education* 17.1, p. 39.

Chan, C.-L. (2011). "Information Security Risk Modeling Using Bayesian Index." In: *The Computer Journal* 54.4, pp. 628–638. DOI: 10.1093/comjnl/bxq059.

Chaparro-Peláez, J., Iglesias-Pradas, S., Rodríguez-Sedano, F. J., and Acquila-Natale, E. (2020). "Extraction, Processing and Visualization of Peer Assessment Data in Moodle." In: *Applied Sciences* 10.1, p. 163. DOI: 10.3390/app10010163.

Chejara, P., Prieto, L. P., Ruiz-Calleja, A., Rodríguez-Triana, M. J., Shankar, S. K., and Kasepalu, R. (2021). "EFAR-MMLA: An Evaluation Framework to Assess and Report Generalizability of Machine Learning Models in MMLA." In: *Sensors* 21.8, p. 2863.

Chen, F., Cui, Y., Lutsyk-King, A., Gao, Y., Liu, X., Cutumisu, M., and Leighton, J. P. (2023). "Validating a Novel Digital Performance-Based Assessment of Data Literacy: Psychometric and Eye-Tracking Analyses." In: *Education and Information Technologies*, pp. 1–28. DOI: 10.1007/s10639-023-12177-7.

Chen, M.-K. and Wang, S.-C. (2010). "A Hybrid Delphi-Bayesian Method to Establish Business Data Integrity Policy: A Benchmark Data Center Case Study." In: *Kybernetes* 39.5. Ed. by D. Dash Wu, pp. 800–824. DOI: 10.1108/03684921011043260.

Chen, T. (2022). "An Argument-Based Validation of an Asynchronous Written Interaction Task." In: *Frontiers in Psychology* 13, pp. 1–10. DOI: 10.3389/fpsyg.2022.889488.

Cherdantseva, Y., Burnap, P., Blyth, A., Eden, P., Jones, K., Soulsby, H., and Stoddart, K. (2016). "A Review of Cyber Security Risk Assessment Methods for SCADA Systems." In: *Computers & Security* 56, pp. 1–27. DOI: 10.1016/j.cose.2015.09.009.

Cho, J.-H., Xu, S., Hurley, P. M., Mackay, M., Benjamin, T., and Beaumont, M. (2019). "STRAM: Measuring the Trustworthiness of Computer-Based Systems." In: *ACM Computing Surveys* 51.6, 128:1–128:47. DOI: 10.1145/3277666.

Choi, Y., Lee, Y., Cho, J., Baek, J., Kim, B., Cha, Y., Shin, D., Bae, C., and Heo, J. (2020). "Towards an Appropriate Query, Key, and Value Computation for Knowledge Tracing." In: *Proceedings of the Seventh ACM Conference on Learning @ Scale*. L@S '20. Online: ACM, pp. 341–344. DOI: 10.1145/3386527.3405945.

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., and Heo, J. (2020). "EdNet: A Large-Scale Hierarchical Dataset in Education." In: *Artificial Intelligence in Education*. Ed. by I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán. AIED '20. Ifrane, Morocco: Springer International Publishing, pp. 69–73. DOI: 10.1007/978-3-030-52240-7_13.

Cholez, H. and Girard, F. (2014). "Maturity Assessment and Process Improvement for Information Security Management in Small and Medium Enterprises." In: *Journal of Software: Evolution and Process* 26.5, pp. 496–503. DOI: 10.1002/smr.1609.

Clauser, B. E., Kane, M. T., and Swanson, D. B. (2002). "Validity Issues for Performance-Based Tests Scored With Computer-Automated Scoring Sys-

tems." In: *Applied Measurement in Education* 15.4, pp. 413–432. DOI: 10.1207/S15324818AME1504_05.

Clunie, L., Morris, N. P., Joynes, V. C., and Pickering, J. D. (2018). "How Comprehensive Are Research Studies Investigating the Efficacy of Technology-Enhanced Learning Resources in Anatomy Education? A Systematic Review." In: *Anatomical Sciences Education* 11.3, pp. 303–319. DOI: 10.1002/ase.1762.

Connolly, J. L., Davidson, M. S., Richard, M., and Skorupka, D. C. W. (2012). *The Trusted Automated eXchange of Indicator Information (TAXII)*. Technical Paper. Mitre Corporation.

Consoli, T., Désiron, J., and Cattaneo, A. (2023). "What Is "Technology Integration" and How Is It Measured in K-12 Education? A Systematic Review of Survey Instruments from 2010 to 2021." In: *Computers & Education* 197, pp. 1–19. DOI: 10.1016/j.compedu.2023.104742.

Cook, D. A., Brydges, R., Ginsburg, S., and Hatala, R. (2015). "A Contemporary Approach to Validity Arguments: A Practical Guide to Kane's Framework." In: *Medical Education* 49.6, pp. 560–575. DOI: 10.1111/medu.12678.

Cormack, G. V. and Grossman, M. R. (2016). "Engineering Quality and Reliability in Technology-Assisted Review." In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. New York, NY, USA: Association for Computing Machinery, pp. 75–84. DOI: 10.1145/2911451.2911510.

Cox, L. A. (2008). "Some Limitations of "Risk = Threat × Vulnerability × Consequence" for Risk Analysis of Terrorist Attacks." In: *Risk Analysis* 28.6, pp. 1749–1761. DOI: 10.1111/j.1539-6924.2008.01142.x.

Cronbach, L. J. and Meehl, P. E. (1955). "Construct Validity in Psychological Tests." In: *Psychological Bulletin* 52.4, pp. 281–302. DOI: 10.1037/h0040957.

da Silva, F. Q. B., Santos, A. L. M., Soares, S., França, A. C. C., Monteiro, C. V. F., and Maciel, F. F. (2011). "Six Years of Systematic Literature Reviews in Software Engineering: An Updated Tertiary Study." In: *Information and Software Technology*. Studying Work Practices in Global Software Engineering 53.9, pp. 899–913. DOI: 10.1016/j.infsof.2011.04.004.

da Silva, F. L., Slodkowski, B. K., da Silva, K. K. A., and Cazella, S. C. (2023). "A Systematic Literature Review on Educational Recommender Systems for Teaching and Learning: Research Trends, Limitations and Opportunities." In: *Education and Information Technologies* 28.3, pp. 3289–3328. DOI: 10.1007/s10639-022-11341-9.

Da Veiga, A. (2018). "An Approach to Information Security Culture Change Combining ADKAR and the ISCA Questionnaire to Aid Transition to the Desired Culture." In: *Information & Computer Security* 26.5, pp. 584–612. DOI: 10.1108/ICS-08-2017-0056.

Da Veiga, A., Astakhova, L. V., Botha, A., and Herselman, M. (2020). "Defining Organisational Information Security Culture—Perspectives from Academia

and Industry." In: *Computers & Security* 92, p. 101713. DOI: 10.1016/j.cose.2020.101713.

Damenu, T. K. and Beaumont, C. (2017). "Analysing Information Security in a Bank Using Soft Systems Methodology." In: *Information & Computer Security* 25.3, pp. 240–258. DOI: 10.1108/ICS-07-2016-0053.

Dantu, R. and Kolan, P. (2005). "Risk Management Using Behavior Based Bayesian Networks." In: *Intelligence and Security Informatics*. Ed. by P. Kantor, G. Muresan, F. Roberts, D. D. Zeng, F.-Y. Wang, H. Chen, and R. C. Merkle. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 115–126. DOI: 10.1007/11427995_10.

Dantu, R., Kolan, P., and Cangussu, J. (2009). "Network Risk Management Using Attacker Profiling." In: *Security and Communication Networks* 2.1, pp. 83–96. DOI: 10.1002/sec.58.

Davis, M. C., Challenger, R., Jayewardene, D. N. W., and Clegg, C. W. (2014). "Advancing Socio-Technical Systems Thinking: A Call for Bravery." In: *Applied Ergonomics*. Advances in Socio-Technical Systems Understanding and Design: A Festschrift in Honour of K.D. Eason 45.2, Part A, pp. 171–180. DOI: 10.1016/j.apergo.2013.02.009.

de Fuentes, J. M., González-Manzano, L., Tapiador, J., and Peris-Lopez, P. (2017). "PRACIS: Privacy-preserving and Aggregatable Cybersecurity Information Sharing." In: *Computers & Security*. Security Data Science and Cyber Threat Management 69, pp. 127–141. DOI: 10.1016/j.cose.2016.12.011.

de las Cuevas, P., Mora, A. M., Merelo, J. J., Castillo, P. A., García-Sánchez, P., and Fernández-Ares, A. (2015). "Corporate Security Solutions for BYOD: A Novel User-Centric and Self-Adaptive System." In: *Computer Communications*. Security and Privacy in Unified Communications\: Challenges and Solutions 68, pp. 83–95. DOI: 10.1016/j.comcom.2015.07.019.

Deci, E. L. and Ryan, R. M. (1985). "The General Causality Orientations Scale: Self-determination in Personality." In: *Journal of Research in Personality* 19.2, pp. 109–134. DOI: 10.1016/0092-6566(85)90023-6.

Deng, M., Wuyts, K., Scandariato, R., Preneel, B., and Joosen, W. (2011). "A Privacy Threat Analysis Framework: Supporting the Elicitation and Fulfillment of Privacy Requirements." In: *Requirements Engineering* 16.1, pp. 3–32. DOI: 10.1007/s00766-010-0115-7.

Dewey, J. (1931). *Philosophy and Civilization*. New York, NY, USA: Minton, Balch & Company.

– (1938). *Logic: The Theory of Inquiry*. New York, NY, USA: Henry Holt & Company.

Dewey, J. and Bentley, A. F. (1949). *Knowing and the Known*. Boston, MA, USA: Beacon Press.

Douglas, K. A., Merzdorf, H. E., Hicks, N. M., Sarfraz, M. I., and Bermel, P. (2020). "Challenges to Assessing Motivation in MOOC Learners: An Application of an Argument-Based Approach." In: *Computers & Education* 150, pp. 1–16. DOI: 10.1016/j.compedu.2020.103829.

Dourado, R. A., Rodrigues, R. L., Ferreira, N., Mello, R. F., Gomes, A. S., and Verbert, K. (2021). "A Teacher-facing Learning Analytics Dashboard for Process-oriented Feedback in Online Learning." In: *Proceedings of the 11th International Learning Analytics and Knowledge Conference*. LAK'21. Irvine, CA, USA: ACM, pp. 482–489. DOI: 10.1145/3448139.3448187.

Drachsler, H. and Greller, W. (2016). "Privacy and Analytics: It's a DELICATE Issue a Checklist for Trusted Learning Analytics." In: *Proceedings of the 6th International Learning Analytics & Knowledge Conference*. LAK '16. Edinburgh, United Kingdom: ACM, pp. 89–98. DOI: 10.1145/2883851.2883893.

Dybå, T. and Dingsøyr, T. (2008). "Strength of Evidence in Systematic Reviews in Software Engineering." In: *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. ESEM '08. New York, NY, USA: Association for Computing Machinery, pp. 178–187. DOI: 10.1145/1414004.1414034.

Eckhart, M., Brenner, B., Ekelhart, A., and Weippl, E. (2019). "Quantitative Security Risk Assessment for Industrial Control Systems: Research Opportunities and Challenges." In: *J. Internet Serv. Inf. Secur.* DOI: 10.22667/JISIS.2019.08.31.052.

Effenberger, T. and Pelánek, R. (2021). "Validity and Reliability of Student Models for Problem-Solving Activities." In: *Proceedings of the 11th International Learning Analytics and Knowledge Conference*. LAK'21. Irvine, CA, USA: ACM, pp. 1–11. DOI: 10.1145/3448139.3448140.

Ekuban, A. and Domingue, J. (2023). "Towards Decentralised Learning Analytics (Positioning Paper)." In: *Companion Proceedings of the ACM Web Conference 2023*. WWW '23. Austin, TX, USA: ACM, pp. 1435–1438. DOI: 10.1145/3543873.3587644.

ENISA (2007). *A Simplified Approach to Risk Management for SMEs*. Report. ENISA.

– (2016). *ENISA Threat Taxonomy*. https://www.enisa.europa.eu/topics/threat-risk-management/threats-and-trends/enisa-threat-landscape/threat-taxonomy.

– (2019). *ENISA Threat Landscape Report 2018*. Report. ENISA.

– (2020). *ENISA Threat Landscape 2020 - List of Top 15 Threats*. Report. ENISA.

Erdt, M., Fernández, A., and Rensing, C. (2015). "Evaluating Recommender Systems for Technology Enhanced Learning: A Quantitative Survey." In: *IEEE Transactions on Learning Technologies* 8.4, pp. 326–344. DOI: 10.1109/TLT.2015.2438867.

Erduran, S., Simon, S., and Osborne, J. (2004). "TAPping into Argumentation: Developments in the Application of Toulmin's Argument Pattern for Studying Science Discourse." In: *Science Education* 88.6, pp. 915–933. DOI: 10.1002/sce.20012.

Eskes, P., Spruit, M., Brinkkemper, S., Vorstman, J., and Kas, M. J. (2016). "The Sociability Score: App-based Social Profiling from a Healthcare Perspec-

tive." In: *Computers in Human Behavior* 59, pp. 39–48. DOI: 10.1016/j.chb.2016.01.024.

European Commission (2016). *SME Definition*. https://ec.europa.eu/growth/smes/sme-definition. Text.

European DIGITAL SME Alliance (2020). *The EU Cybersecurity Act and the Role of Standards for SMEs - Position Paper*. Tech. rep. Brussels: European DIGITAL SME Alliance.

Evesti, A. and Ovaska, E. (2013). "Comparison of Adaptive Information Security Approaches." In: *ISRN Artificial Intelligence*.

Ezhei, M. and Tork Ladani, B. (2017). "Information Sharing vs. Privacy: A Game Theoretic Analysis." In: *Expert Systems with Applications* 88, pp. 327–337. DOI: 10.1016/j.eswa.2017.06.042.

Fachola, C., Tornaría, A., Bermolen, P., Capdehourat, G., Etcheverry, L., and Fariello, M. I. (2023). "Federated Learning for Data Analytics in Education." In: *Data* 8.2, p. 43. DOI: 10.3390/data8020043.

Faiella, M., Gonzalez-Granadillo, G., Medeiros, I., Azevedo, R., and Gonzalez-Zarzosa, S. (2021). "Enriching Threat Intelligence Platforms Capabilities." In: *Proceedings of the 16th International Joint Conference on E-Business and Telecommunications - SECRYPT*. Prague, Czech Republic: SciTePress - Science and and Technology Publications, pp. 37–48.

Fan, Y., Lim, L., van der Graaf, J., Kilgour, J., Raković, M., Moore, J., Molenaar, I., Bannert, M., and Gašević, D. (2022). "Improving the Measurement of Self-Regulated Learning Using Multi-Channel Data." In: *Metacognition and Learning* 17.3, pp. 1025–1055.

Fan, Y., van der Graaf, J., Lim, L., Raković, M., Singh, S., Kilgour, J., Moore, J., Molenaar, I., Bannert, M., and Gašević, D. (2022). "Towards Investigating the Validity of Measurement of Self-Regulated Learning Based on Trace Data." In: *Metacognition and Learning* 17.3, pp. 949–987. DOI: 10.1007/s11409-022-09291-1.

Feng, N., Wang, H. J., and Li, M. (2014). "A Security Risk Analysis Model for Information Systems: Causal Relationships of Risk Factors and Vulnerability Propagation Analysis." In: *Information Sciences*. Business Intelligence in Risk Management 256, pp. 57–73. DOI: 10.1016/j.ins.2013.02.036.

Feng, W., Tang, J., and Liu, T. X. (2019). "Understanding Dropouts in MOOCs." In: *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. Vol. 33. AAAI '19. Honolulu, HI, USA: PKP Publishing Services, pp. 517–524. DOI: 10.1609/aaai.v33i01.3301517.

Ferguson, N. and Schneier, B. (2003). *Practical Cryptography*. Wiley.

Ferguson, R., Clow, D., Griffiths, D., and Brasher, A. (2019). "Moving Forward with Learning Analytics: Expert Views." In: *Journal of Learning Analytics* 6.3, pp. 43–59. DOI: 10.18608/jla.2019.63.8.

Fincham, E., Whitelock-Wainwright, A., Kovanović, V., Joksimović, S., van Staalduinen, J.-P., and Gašević, D. (2019). "Counting Clicks Is Not Enough: Validating a Theorized Model of Engagement in Learning Analytics." In:

Proceedings of the 9th International Learning Analytics and Knowledge Conference. LAK'19. Tempe, AZ, USA: ACM, pp. 501–510. DOI: 10.1145/3303772.3303775.

Galaige, J., Torrisi-Steele, G., Binnewies, S., and Wang, K. (2018). "What Is Important in Student-Facing Learning Analytics? A User-Centered Design Approach." In: *Proceedings of the 22nd Pacific Asia Conference on Information Systems*. PACIS'18. Yokohoma, Japan: AIS, pp. 1248–1261.

Gañán, D., Caballé, S., Clarisó, R., Conesa, J., and Bañeres, D. (2017). "ICT-FLAG: A Web-Based e-Assessment Platform Featuring Learning Analytics and Gamification." In: *International Journal of Web Information Systems* 13.1, pp. 25–54.

Gardner, J., Yu, R., Nguyen, Q., Brooks, C., and Kizilcec, R. (2023). "Cross-Institutional Transfer Learning for Educational Models: Implications for Model Performance, Fairness, and Equity." In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. Chicago, IL, USA: ACM, pp. 1664–1684. DOI: 10.1145/3593013.3594107.

Gašević, D., Dawson, S., Rogers, T., and Gasevic, D. (2016). "Learning Analytics Should Not Promote One Size Fits All: The Effects of Instructional Conditions in Predicting Academic Success." In: *The Internet and Higher Education* 28, pp. 68–84.

Gašević, D., Greiff, S., and Shaffer, D. W. (2022). "Towards Strengthening Links between Learning Analytics and Assessment: Challenges and Potentials of a Promising New Bond." In: *Computers in Human Behavior* 134, pp. 1–7. DOI: 10.1016/j.chb.2022.107304.

Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., and Hartling, L. (2019). "Performance and Usability of Machine Learning for Screening in Systematic Reviews: A Comparative Evaluation of Three Tools." In: *Systematic Reviews* 8.1, p. 278. DOI: 10.1186/s13643-019-1222-2.

Geertz, C. (1973). "Thick Description: Toward an Interpretive Theory of Culture." In: *The Interpretation Of Cultures*. New York, NY, USA: Basic Books, pp. 3–30.

GEIGER Consortium (2020). *GEIGER Project Website*. https://project.cyber-geiger.eu/.

Giannakos, M. N., Chorianopoulos, K., and Chrisochoides, N. (2015). "Making Sense of Video Analytics: Lessons Learned from Clickstream Interactions, Attitudes, and Learning Outcome in a Video-Assisted Course." In: *International Review of Research in Open and Distributed Learning* 16.1, pp. 260–283. DOI: 10.19173/irrodl.v16i1.1976.

Glaser, B. G. and Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. 1st. London, UK: Aldine Publishing Company.

Glass, G. (1976). "Primary, Secondary, and Meta-Analysis of Research." In: *Educational Researcher* 5.10, pp. 3–8. DOI: 10.3102/0013189X005010003.

Goldhammer, F., Hahnel, C., Kroehne, U., and Zehner, F. (2021). "From Byproduct to Design Factor: On Validating the Interpretation of Process Indicators

Based on Log Data." In: *Large-scale Assessments in Education* 9.1, pp. 1–25. DOI: 10.1186/s40536-021-00113-5.

Goldkuhl, G. (2004). "Design Theories in Information Systems - A Need for Multi-Grounding." In: *Journal of Information Technology Theory and Application (JITTA)* 6.2, pp. 59–72.

Goldkuhl, G. and Cronholm, S. (2010). "Adding Theoretical Grounding to Grounded Theory: Toward Multi-Grounded Theory." In: *International Journal of Qualitative Methods* 9.2, pp. 187–205. DOI: 10.1177/160940691000900205.

Goldkuhl, G., Cronholm, S., and Lind, M. (2020). "Multi-Grounded Action Research." In: *Information Systems and e-Business Management* 18.2, pp. 121–156. DOI: 10.1007/s10257-020-00469-1.

Gollmann, D., Herley, C., Koenig, V., Pieters, W., and Sasse, M. A. (2015). "Socio-Technical Security Metrics (Dagstuhl Seminar 14491)." In: *Dagstuhl Reports* 4.12, pp. 1–28. DOI: 10.4230/DagRep.4.12.1.

Gonzalez-Granadillo, G., Faiella, M., Medeiros, I., Azevedo, R., and Gonzalez-Zarzosa, S. (2019). "Enhancing Information Sharing and Visualization Capabilities in Security Data Analytic Platforms." In: *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. Portland, OR, USA: IEEE, pp. 1–8. DOI: 10.1109/DSN-W.2019.00009.

Gough, D., Oliver, S., and Thomas, J. (2017). *An Introduction to Systematic Reviews*. SAGE.

Graf, R. and King, R. (2018). "Neural Network and Blockchain Based Technique for Cyber Threat Intelligence and Situational Awareness." In: *2018 10th International Conference on Cyber Conflict (CyCon)*. Tallinn, Estonia: IEEE, pp. 409–426. DOI: 10.23919/CYCON.2018.8405028.

Gratian, M., Bandi, S., Cukier, M., Dykstra, J., and Ginther, A. (2018). "Correlating Human Traits and Cyber Security Behavior Intentions." In: *Computers & Security* 73, pp. 345–358. DOI: 10.1016/j.cose.2017.11.015.

Guba, E. G. (1981). "Criteria for Assessing the Trustworthiness of Naturalistic Inquiries." In: *ECTJ* 29.2, pp. 75–91. DOI: 10.1007/BF02766777.

Guo, S. and Zeng, D. (2020). "Pedagogical Data Federation toward Education 4.0." In: *Proceedings of the 6th International Conference on Frontiers of Educational Technologies*. ICFET '20. Tokyo, Japan: ACM, pp. 51–55. DOI: 10.1145/3404709.3404751.

Gursoy, M. E., Inan, A., Nergiz, M. E., and Saygin, Y. (2017). "Privacy-Preserving Learning Analytics: Challenges and Techniques." In: *IEEE Transactions on Learning Technologies* 10.1, pp. 68–81. DOI: 10.1109/TLT.2016.2607747.

Gusenbauer, M. and Haddaway, N. R. (2020). "Which Academic Search Systems Are Suitable for Systematic Reviews or Meta-Analyses? Evaluating Retrieval Qualities of Google Scholar, PubMed, and 26 Other Resources." In: *Research Synthesis Methods* 11.2, pp. 181–217. DOI: 10.1002/jrsm.1378.

Hall, T., Beecham, S., Bowes, D., Gray, D., and Counsell, S. (2012). "A Systematic Literature Review on Fault Prediction Performance in Software

Engineering." In: *IEEE Transactions on Software Engineering* 38.6, pp. 1276–1304. DOI: 10.1109/TSE.2011.103.

Halvorsrud, R., Kvale, K., and Følstad, A. (2016). "Improving Service Quality through Customer Journey Analysis." In: *Journal of Service Theory and Practice* 26.6, pp. 840–867. DOI: 10.1108/JSTP-05-2015-0111.

Hanus, B. and Wu, Y. " (2016). "Impact of Users' Security Awareness on Desktop Security Behavior: A Protection Motivation Theory Perspective." In: *Information Systems Management* 33.1, pp. 2–16. DOI: 10.1080/10580530.2015.1117842.

Harrison, H., Griffin, S. J., Kuhn, I., and Usher-Smith, J. A. (2020). "Software Tools to Support Title and Abstract Screening for Systematic Reviews in Healthcare: An Evaluation." In: *BMC Medical Research Methodology* 20.1, p. 7. DOI: 10.1186/s12874-020-0897-3.

He, S., Fu, J., Jiang, W., Cheng, Y., Chen, J., and Guo, Z. (2020). "BloTISRT: Blockchain-based Threat Intelligence Sharing and Rating Technology." In: *Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies*. CIAT 2020. New York, NY, USA: Association for Computing Machinery, pp. 524–534. DOI: 10.1145/3444370.3444623.

He, S., Lee, G. M., Han, S., and Whinston, A. B. (2016). "How Would Information Disclosure Influence Organizations' Outbound Spam Volume? Evidence from a Field Experiment." In: *Journal of Cybersecurity* 2.1, pp. 99–118. DOI: 10.1093/cybsec/tyw011.

He, W., Li, H., and Li, J. (2019). "Unknown Vulnerability Risk Assessment Based on Directed Graph Models: A Survey." In: *IEEE Access* 7, pp. 168201–168225. DOI: 10.1109/ACCESS.2019.2954092.

Heidt, M., Gerlach, J. P., and Buxmann, P. (2019). "Investigating the Security Divide between SME and Large Companies: How SME Characteristics Influence Organizational IT Security Investments." In: *Information Systems Frontiers* 21.6, pp. 1285–1305. DOI: 10.1007/s10796-019-09959-1.

Heil, J. and Ifenthaler, D. (2023). "Online Assessment in Higher Education: A Systematic Review." In: *Online Learning* 27.1, pp. 187–218. DOI: 10.24059/olj.v27i1.3398.

Herath, T. and Rao, H. R. (2009). "Encouraging Information Security Behaviors in Organizations: Role of Penalties, Pressures and Perceived Effectiveness." In: *Decision Support Systems* 47.2, pp. 154–165. DOI: 10.1016/j.dss.2009.02.005.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). "Design Science in Information Systems Research." In: *MIS Quarterly* 28.1, pp. 75–105.

Hiatt, J. (2006). *ADKAR: A Model for Change in Business, Government, and Our Community*. Prosci.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A. (2019). *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons.

Hlosta, M., Zdrahal, Z., and Zendulka, J. (2017). "Ouroboros: Early Identification of at-Risk Students without Models Based on Legacy Data." In: *Proceedings of the 7th International Learning Analytics & Knowledge Conference*. LAK '17. Vancouver, BC, Canada: ACM, pp. 6–15. DOI: 10.1145/3027385.3027449.

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., and Veldkamp, B. P. (2019). "A General Framework for the Validation of Embedded Formative Assessment." In: *Journal of Educational Measurement* 56.4, pp. 715–732. DOI: 10.1111/jedm.12234.

Howell, J. A., Roberts, L. D., and Mancini, V. O. (2018). "Learning Analytics Messages: Impact of Grade, Sender, Comparative Information and Message Style on Student Affect and Academic Resilience." In: *Computers in Human Behavior* 89, pp. 8–15. DOI: 10.1016/j.chb.2018.07.021.

Huang, H., Gao, Y., Yan, M., and Zhang, X. (2020). "Research on Industrial Internet Security Emergency Management Framework Based on Blockchain: Take China as an Example." In: *CNCERT 2020: Cyber Security*. Ed. by W. Lu, Q. Wen, Y. Zhang, B. Lang, W. Wen, H. Yan, C. Li, L. Ding, R. Li, and Y. Zhou. Communications in Computer and Information Science. Beijing, China: Springer, pp. 71–85. DOI: 10.1007/978-981-33-4922-3_6.

Huggins-Manley, A. C., Booth, B. M., and D'Mello, S. K. (2022). "Toward Argument-Based Fairness with an Application to AI-Enhanced Educational Assessments." In: *Journal of Educational Measurement* 59.3, pp. 362–388. DOI: 10.1111/jedm.12334.

Husák, M., Bajtoš, T., Kašpar, J., Bou-Harb, E., and Čeleda, P. (2020). "Predictive Cyber Situational Awareness and Personalized Blacklisting: A Sequential Rule Mining Approach." In: *ACM Transactions on Management Information Systems* 11.4, 19:1–19:16. DOI: 10.1145/3386250.

Husák, M., Bartoš, V., Sokol, P., and Gajdoš, A. (2021). "Predictive Methods in Cyber Defense: Current Experience and Research Challenges." In: *Future Generation Computer Systems* 115, pp. 517–530. DOI: 10.1016/j.future.2020.10.006.

Husák, M., Komárková, J., Bou-Harb, E., and Čeleda, P. (2019). "Survey of Attack Projection, Prediction, and Forecasting in Cyber Security." In: *IEEE Communications Surveys & Tutorials* 21.1, pp. 640–660. DOI: 10.1109/COMST.2018.2871866.

Husari, G., Niu, X., Chu, B., and Al-Shaer, E. (2018). "Using Entropy and Mutual Information to Extract Threat Actions from Cyber Threat Intelligence." In: *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*. Miami, FL, USA: IEEE, pp. 1–6. DOI: 10.1109/ISI.2018.8587343.

Iannacone, M. D. and Bridges, R. A. (2020). "Quantifiable & Comparable Evaluations of Cyber Defensive Capabilities: A Survey & Novel, Unified Approach." In: *Computers & Security* 96, p. 101907. DOI: 10.1016/j.cose.2020.101907.

Ifenthaler, D. and Widanapathirana, C. (2014). "Development and Validation of a Learning Analytics Framework: Two Case Studies Using Support Vector Machines." In: *Technology, Knowledge and Learning* 19.1, pp. 221–240. DOI: 10.1007/s10758-014-9226-4.

International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) (2012). *ISO/IEC 27032:2012 Information Technology — Security Techniques — Guidelines for Cybersecurity*. Tech. rep. ISO/IEC.

– (2013). *ISO/IEC 27002:2013 Information Technology — Security Techniques — Code of Practice for Information Security Controls*. Tech. rep. ISO/IEC.

Jantsch, E. (1970). "Inter- and Transdisciplinary University: A Systems Approach to Education and Innovation." In: *Policy Sciences* 1.1, pp. 403–428. DOI: 10.1007/BF00145222.

Jaquith, A. (2007). *Security Metrics: Replacing Fear, Uncertainty, and Doubt*. Addison-Wesley.

Jeng, T.-H., Chan, W.-M., Luo, W.-Y., Huang, C.-C., Chen, C.-C., and Chen, Y.-M. (2019). "NetFlowTotal: A Cloud Service Integration Platform for Malicious Traffic Analysis and Collaboration." In: *Proceedings of the 2nd International Conference on Computing and Big Data*. ICCBD 2019. New York, NY, USA: Association for Computing Machinery, pp. 154–160. DOI: 10.1145/3366650.3366669.

Jing, X., Yan, Z., and Pedrycz, W. (2019). "Security Data Collection and Data Analytics in the Internet: A Survey." In: *IEEE Communications Surveys & Tutorials* 21.1, pp. 586–618. DOI: 10.1109/COMST.2018.2863942.

Jo, I.-H., Kim, D., and Yoon, M. (2014). "Analyzing the Log Patterns of Adult Learners in LMS Using Learning Analytics." In: *Proceedings of the 4th International Learning Analytics and Knowledge Conference*. LAK '14. Indianapolis, IN, USA: ACM, pp. 183–187.

Johnson, C., Badger, M., Waltermire, D., Snyder, J., and Skorupka, C. (2016). *Guide to Cyber Threat Information Sharing*. Technical Report NIST Special Publication (SP) 800-150. National Institute of Standards and Technology. DOI: 10.6028/NIST.SP.800-150.

Jones, K. M. L., Rubel, A., and LeClere, E. (2020). "A Matter of Trust: Higher Education Institutions as Information Fiduciaries in an Age of Educational Data Mining and Learning Analytics." In: *Journal of the Association for Information Science and Technology* 71.10, pp. 1227–1241. DOI: 10.1002/asi.24327.

Kaila, U. and Nyman, L. (2018). "Information Security Best Practices: First Steps for Startups and SMEs." In: *Technology Innovation Management Review* 8.11, pp. 32–42. DOI: 10.22215/timreview/1198.

Kam, H.-J., Menard, P., Ormond, D., and Crossler, R. E. (2020). "Cultivating Cybersecurity Learning: An Integration of Self-Determination and Flow." In: *Computers & Security* 96, p. 101875. DOI: 10.1016/j.cose.2020.101875.

Kampanakis, P. (2014). "Security Automation and Threat Information-Sharing Options." In: *IEEE Security Privacy* 12.5, pp. 42–51. DOI: 10.1109/MSP.2014.99.

Kane, M. T. (1992). "An Argument-Based Approach to Validity." In: *Psychological Bulletin* 112.3, pp. 527–535. DOI: 10.1037/0033-2909.112.3.527.

– (2004). "Certification Testing as an Illustration of Argument-Based Validation." In: *Measurement: Interdisciplinary Research and Perspectives* 2.3, pp. 135–170. DOI: 10.1207/s15366359mea0203_1.

– (2013a). "The Argument-Based Approach to Validation." In: *School Psychology Review* 42.4. Ed. by M. Burns, pp. 448–457. DOI: 10.1080/02796015.2013.12087465.

– (2013b). "Validating the Interpretations and Uses of Test Scores." In: *Journal of Educational Measurement* 50.1, pp. 1–73. DOI: 10.1111/jedm.12000.

Karlsson, F. and Ågerfalk, P. J. (2007). "Multi-Grounded Action Research in Method Engineering: The MMC Case." In: *Situational Method Engineering: Fundamentals and Experiences*. Ed. by J. Ralyté, S. Brinkkemper, and B. Henderson-Sellers. IFIP — The International Federation for Information Processing. Geneva, Switzerland: Springer, pp. 19–32. DOI: 10.1007/978-0-387-73947-2_4.

Kärner, T., Warwas, J., and Schumann, S. (2021). "A Learning Analytics Approach to Address Heterogeneity in the Classroom: The Teachers' Diagnostic Support System." In: *Technology, Knowledge and Learning* 26.1, pp. 31–52. DOI: 10.1007/s10758-020-09448-4.

Khan, M. U., Sherin, S., Iqbal, M. Z., and Zahid, R. (2019). "Landscaping Systematic Mapping Studies in Software Engineering: A Tertiary Study." In: *Journal of Systems and Software* 149, pp. 396–436. DOI: 10.1016/j.jss.2018.12.018.

Khramtsova, E., Hammerschmidt, C., Lagraa, S., and State, R. (2020). "Federated Learning For Cyber Security: SOC Collaboration For Malicious URL Detection." In: *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. Singapore, Singapore: IEEE, pp. 1316–1321. DOI: 10.1109/ICDCS47774.2020.00171.

Kim, D., Park, Y., Yoon, M., and Jo, I.-H. (2016). "Toward Evidence-Based Learning Analytics: Using Proxy Variables to Improve Asynchronous Online Discussion Environments." In: *The Internet and Higher Education* 30, pp. 30–43. DOI: 10.1016/j.iheduc.2016.03.002.

Kim, E., Kim, K., Shin, D., Jin, B., and Kim, H. (2018). "CyTIME: Cyber Threat Intelligence ManagEment Framework for Automatically Generating Security Rules." In: *Proceedings of the 13th International Conference on Future Internet Technologies*. CFI 2018. Seoul, Republic of Korea: Association for Computing Machinery, pp. 1–5. DOI: 10.1145/3226052.3226056.

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., and Lee, D. (2003). "A Taxonomy of Dirty Data." In: *Data Mining and Knowledge Discovery* 7.1, pp. 81–99. DOI: 10.1023/A:1021564703268.

Kitchenham, B. A., Budgen, D., and Brereton, P. (2015). *Evidence-Based Software Engineering and Systematic Reviews*. CRC Press.

Kitchenham, B. A. and Charters, S. (2007). *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report. Keele University and University of Durham.

Kitchenham, B. A., Dyba, T., and Jorgensen, M. (2004). "Evidence-Based Software Engineering." In: *Proceedings. 26th International Conference on Software Engineering*, pp. 273–281. DOI: 10.1109/ICSE.2004.1317449.

Kitchenham, B. A., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., Emam, K. E., and Rosenberg, J. (2002). "Preliminary Guidelines for Empirical Research in Software Engineering." In: *IEEE Transactions on Software Engineering* 28.8, pp. 721–734. DOI: 10.1109/TSE.2002.1027796.

Kitchenham, B. A., Pretorius, R., Budgen, D., Pearl Brereton, O., Turner, M., Niazi, M., and Linkman, S. (2010). "Systematic Literature Reviews in Software Engineering – A Tertiary Study." In: *Information and Software Technology* 52.8, pp. 792–805. DOI: 10.1016/j.infsof.2010.03.006.

Kitto, K., Buckingham Shum, S., and Gibson, A. (2018). "Embracing Imperfection in Learning Analytics." In: *Proceedings of the 8th International Learning Analytics and Knowledge Conference*. LAK '18. Sydney, Australia: ACM, pp. 451–460. DOI: 10.1145/3170358.3170413.

Kizilcec, R. F., Pérez-Sanagustín, M., and Maldonado, J. J. (2017). "Self-Regulated Learning Strategies Predict Learner Behavior and Goal Attainment in Massive Open Online Courses." In: *Computers & Education* 104, pp. 18–33. DOI: 10.1016/j.compedu.2016.10.001.

Knight, S., Shum, S. B., and Littleton, K. (2014). "Epistemology, Assessment, Pedagogy: Where Learning Meets Analytics in the Middle Space." In: *Journal of Learning Analytics* 1.2, pp. 23-47-23–47. DOI: 10.18608/jla.2014.12.3.

Knowles, W., Prince, D., Hutchison, D., Disso, J. F. P., and Jones, K. (2015). "A Survey of Cyber Security Management in Industrial Control Systems." In: *International Journal of Critical Infrastructure Protection* 9, pp. 52–80. DOI: 10.1016/j.ijcip.2015.02.002.

Koloveas, P., Chantzios, T., Alevizopoulou, S., Skiadopoulos, S., and Tryfonopoulos, C. (2021). "inTIME: A Machine Learning-Based Framework for Gathering and Leveraging Web Data to Cyber-Threat Intelligence." In: *Electronics* 10.7, p. 818. DOI: 10.3390/electronics10070818.

Kordy, B., Pièvre-Cambacédès, L., and Schweitzer, P. (2014). "DAG-based Attack and Defense Modeling: Don't Miss the Forest for the Attack Trees." In: *Computer Science Review* 13–14, pp. 1–38. DOI: 10.1016/j.cosrev.2014.07.001.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Kure, H. and Islam, S. (2019). "Cyber Threat Intelligence for Improving Cybersecurity and Risk Management in Critical Infrastructure." In: *JUCS* -

Journal of Universal Computer Science 25(11), pp. 1478–1502. DOI: 10.3217/jucs-025-11-1478.

Kuzilek, J., Hlosta, M., and Zdrahal, Z. (2017). "Open University Learning Analytics Dataset." In: *Scientific Data* 4.1, p. 170171. DOI: 10.1038/sdata.2017.171.

Lai, J. W. M. and Bower, M. (2019). "How Is the Use of Technology in Education Evaluated? A Systematic Review." In: *Computers & Education* 133, pp. 27–42. DOI: 10.1016/j.compedu.2019.01.010.

– (2020). "Evaluation of Technology Use in Education: Findings from a Critical Analysis of Systematic Literature Reviews." In: *Journal of Computer Assisted Learning* 36.3, pp. 241–259. DOI: 10.1111/jcal.12412.

Lai, J. W. M., Bower, M., De Nobile, J., and Breyer, Y. (2022). "What Should We Evaluate When We Use Technology in Education?" In: *Journal of Computer Assisted Learning* 38.3, pp. 743–757. DOI: 10.1111/jcal.12645.

Law, E. L.-C. and Heintz, M. (2021). "Augmented Reality Applications for K-12 Education: A Systematic Review from the Usability and User Experience Perspective." In: *International Journal of Child-Computer Interaction* 30, pp. 1–23. DOI: 10.1016/j.ijcci.2021.100321.

Lawrence, M. G., Williams, S., Nanz, P., and Renn, O. (2022). "Characteristics, Potentials, and Challenges of Transdisciplinary Research." In: *One Earth* 5.1, pp. 44–61. DOI: 10.1016/j.oneear.2021.12.010.

Lazarovitz, L. (2021). "Deconstructing the SolarWinds Breach." In: *Computer Fraud & Security* 2021.6, pp. 17–19. DOI: 10.1016/S1361-3723(21)00065-8.

Lee, Y. and Larsen, K. R. (2009). "Threat or Coping Appraisal: Determinants of SMB Executives' Decision to Adopt Anti-Malware Software." In: *European Journal of Information Systems* 18.2, pp. 177–187. DOI: 10.1057/ejis.2009.11.

Lella, I., Theocharidou, M., Tsekmezoglou, E., Malatras, A., Garcia, S., and Valeros, V. (2021). *Threat Landscape for Supply Chain Attacks*. Technical Report. ENISA.

Lemay, A., Calvet, J., Menet, F., and Fernandez, J. M. (2018). "Survey of Publicly Available Reports on Advanced Persistent Threat Actors." In: *Computers & Security* 72, pp. 26–59. DOI: 10.1016/j.cose.2017.08.005.

Leszczyna, R., Wallis, T., and Wróbel, M. R. (2019). "Developing Novel Solutions to Realise the European Energy – Information Sharing & Analysis Centre." In: *Decision Support Systems* 122, p. 113067. DOI: 10.1016/j.dss.2019.05.007.

Leszczyna, R. and Wróbel, M. R. (2019). "Threat Intelligence Platform for the Energy Sector." In: *Software: Practice and Experience* 49.8, pp. 1225–1254. DOI: 10.1002/spe.2705.

Li, W., Gao, M., Li, H., Xiong, Q., Wen, J., and Wu, Z. (2016). "Dropout Prediction in MOOCs Using Behavior Features and Multi-View Semi-Supervised Learning." In: *2016 International Joint Conference on Neural Networks*. IJCNN. Vancouver, BC, Canada: IEEE, pp. 3130–3137. DOI: 10.1109/IJCNN.2016.7727598.

Li, X., Li, H., Sun, B., and Wang, F. (2018). "Assessing Information Security Risk for an Evolving Smart City Based on Fuzzy and Grey FMEA." In: *Journal of Intelligent & Fuzzy Systems* 34, pp. 2491–2501. DOI: 10.3233/JIFS-172097.

Liang, X. and Xiao, Y. (2013). "Game Theory for Network Security." In: *IEEE Communications Surveys & Tutorials* 15.1, pp. 472–486. DOI: 10.1109/SURV.2012.062612.00056.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., and Moher, D. (2009). "The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration." In: *Journal of Clinical Epidemiology* 62.10, e1–e34. DOI: 10.1016/j.jclinepi.2009.06.006.

Lin, Y., Wang, H., Yang, B., Liu, M., Li, Y., and Zhang, Y. (2019). "A Blackboard Sharing Mechanism for Community Cyber Threat Intelligence Based on Multi-Agent System." In: *ML4CS 2019: Machine Learning for Cyber Security*. Ed. by X. Chen, X. Huang, and J. Zhang. Lecture Notes in Computer Science. Xi'an, China: Springer International Publishing, pp. 253–270. DOI: 10.1007/978-3-030-30619-9_18.

Lincoln, Y. S. and Guba, E. G. (1986). "But Is It Rigorous? Trustworthiness and Authenticity in Naturalistic Evaluation." In: *New Directions for Program Evaluation* 1986.30, pp. 73–84. DOI: 10.1002/ev.1427.

Lippmann, R. P. and Riordan, J. F. (2016). "Threat-Based Risk Assessment for Enterprise Networks." In: *Lincoln Laboratory Journal* 22.1, pp. 33–45.

Lippmann, R. P., Riordan, J. F., Yu, T. H., and Watson, K. K. (2012). *Continuous Security Metrics for Prevalent Network Threats: Introduction and First Four Metrics*. Project Report. Lexington, MA, USA: Massachusetts Institute of Technology.

Liu, Q., Geertshuis, S., and Grainger, R. (2020). "Understanding Academics' Adoption of Learning Technologies: A Systematic Review." In: *Computers & Education* 151, pp. 1–19. DOI: 10.1016/j.compedu.2020.103857.

Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., and Liu, M. (2015). "Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents." In: *24th USENIX Security Symposium (USENIX Security 15)*, pp. 1009–1024.

Lo, C.-C. and Chen, W.-J. (2012). "A Hybrid Information Security Risk Assessment Procedure Considering Interdependences between Controls." In: *Expert Systems with Applications* 39.1, pp. 247–257. DOI: 10.1016/j.eswa.2011.07.015.

Long, T., Qin, J., Shen, J., Zhang, W., Xia, W., Tang, R., He, X., and Yu, Y. (2022). "Improving Knowledge Tracing with Collaborative Information." In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. WSDM '22. Online: ACM, pp. 599–607. DOI: 10.1145/3488560.3498374.

Luh, R., Temper, M., Tjoa, S., Schrittwieser, S., and Janicke, H. (2020). "PenQuest: A Gamified Attacker/Defender Meta Model for Cyber Security Assessment and Education." In: *Journal of Computer Virology and Hacking Techniques* 16.1, pp. 19–61. DOI: 10.1007/s11416-019-00342-x.

Malatji, M., Marnewick, A., and von Solms, S. (2020). "Validation of a Socio-Technical Management Process for Optimising Cybersecurity Practices." In: *Computers & Security* 95, p. 101846. DOI: 10.1016/j.cose.2020.101846.

Malatji, M., Von Solms, S., and Marnewick, A. (2019). "Socio-Technical Systems Cybersecurity Framework." In: *Information & Computer Security* 27.2, pp. 233–272. DOI: 10.1108/ICS-03-2018-0031.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., and Munoz-Gama, J. (2018). "Mining Theory-Based Patterns from Big Data: Identifying Self-Regulated Learning Strategies in Massive Open Online Courses." In: *Computers in Human Behavior* 80, pp. 179–196.

Manadhata, P. K. and Wing, J. M. (2011). "An Attack Surface Metric." In: *IEEE Transactions on Software Engineering* 37.3, pp. 371–386. DOI: 10.1109/TSE.2010.60.

Manfredi, S., Ranise, S., Sciarretta, G., and Tomasi, A. (2021). "TLSAssistant Goes FINSEC A Security Platform Integration Extending Threat Intelligence Language." In: *International Workshop on Cyber-Physical Security for Critical Infrastructures Protection (CPS4CIP 2020)*. Ed. by H. Abie, S. Ranise, L. Verderame, E. Cambiaso, R. Ugarelli, G. Giunta, I. Praça, and F. Battisti. Lecture Notes in Computer Science. Guildford, UK: Springer International Publishing, pp. 16–30. DOI: 10.1007/978-3-030-69781-5_2.

Manifavas, C., Fysarakis, K., Rantos, K., and Hatzivasilis, G. (2014). "DSAPE – Dynamic Security Awareness Program Evaluation." In: *Human Aspects of Information Security, Privacy, and Trust*. Ed. by T. Tryfonas and I. Askoxylakis. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 258–269. DOI: 10.1007/978-3-319-07620-1_23.

Marconato, G. V., Kaâniche, M., and Nicomette, V. (2013). "A Vulnerability Life Cycle-Based Security Modeling and Evaluation Approach." In: *The Computer Journal* 56.4, pp. 422–439. DOI: 10.1093/comjnl/bxs112.

Marcos-Pablos, S. and García-Peñalvo, F. J. (2020). "Information Retrieval Methodology for Aiding Scientific Database Search." In: *Soft Computing* 24.8, pp. 5551–5560. DOI: 10.1007/s00500-018-3568-0.

Marcos-Pablos, S. and García-Peñalvo, F. J. (2018). "Decision Support Tools for SLR Search String Construction." In: *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*. TEEM'18. New York, NY, USA: Association for Computing Machinery, pp. 660–667. DOI: 10.1145/3284179.3284292.

Marinos, L. and Sfakianakis, A. (2013). *ENISA Threat Landscape 2012*. Report. ENISA.

Marshall, I. J. and Wallace, B. C. (2019). "Toward Systematic Review Automation: A Practical Guide to Using Machine Learning Tools in Research

Synthesis." In: *Systematic Reviews* 8.1, p. 163. DOI: 10.1186/s13643-019-1074-9.

Martens, M., De Wolf, R., and De Marez, L. (2019). "Investigating and Comparing the Predictors of the Intention towards Taking Security Measures against Malware, Scams and Cybercrime in General." In: *Computers in Human Behavior* 92, pp. 139–150. DOI: 10.1016/j.chb.2018.11.002.

Matcha, W., Gašević, D., Jovanović, J., Uzir, N. A., Oliver, C. W., Murray, A., and Gasevic, D. (2020). "Analytics of Learning Strategies: The Association with the Personality Traits." In: *Proceedings of the 10th International Learning Analytics and Knowledge Conference*. LAK '20. Frankfurt, Germany: ACM, pp. 151–160. DOI: 10.1145/3375462.3375534.

Mavroeidis, V. and Bromander, S. (2017). "Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence." In: *2017 European Intelligence and Security Informatics Conference (EISIC)*. Athens, Greece: IEEE, pp. 91–98. DOI: 10.1109/EISIC.2017.20.

McKeever, P., Allhof, M., Corsi, A., Sowa, I., and Monti, A. (2020). "Wide-Area Cyber-security Analytics Solution for Critical Infrastructures." In: *2020 6th IEEE International Energy Conference (ENERGYCon)*. Gammarth, Tunisia: IEEE, pp. 34–37. DOI: 10.1109/ENERGYCon48941.2020.9236483.

McKenney, S. and Reeves, T. C. (2018). *Conducting Educational Design Research*. 2nd. London, UK: Routledge. DOI: 10.4324/9781315105642.

– (2021). "Educational Design Research: Portraying, Conducting, and Enhancing Productive Scholarship." In: *Medical Education* 55.1, pp. 82–92. DOI: 10.1111/medu.14280.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y (2017). "Communication-Efficient Learning of Deep Networks from Decentralized Data." In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1273–1282.

Menard, P., Bott, G. J., and Crossler, R. E. (2017). "User Motivations in Protecting Information Security: Protection Motivation Theory Versus Self-Determination Theory." In: *Journal of Management Information Systems* 34.4, pp. 1203–1230. DOI: 10.1080/07421222.2017.1394083.

Messick, S. (1989). "Validity." In: *Educational Measurement*. Ed. by R. L. Linn. 3rd ed. The American Council on Education/Macmillan Series on Higher Education. New York, NY, USA: Macmillan Publishing Co, Inc, pp. 13–103.

– (1995). "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." In: *American Psychologist* 50.9, pp. 741–749. DOI: 10.1037/0003-066X.50.9.741.

Mijnhardt, F., Baars, T., and Spruit, M. (2016). "Organizational Characteristics Influencing SME Information Security Maturity." In: *Journal of Computer Information Systems* 56.2, pp. 106–115. DOI: 10.1080/08874417.2016.1117369.

Milligan, S. K. (2018). "Methodological Foundations for the Measurement of Learning in Learning Analytics." In: *Proceedings of the 8th International Learning Analytics and Knowledge Conference*. LAK '18. Sydney, Australia: ACM, pp. 466–470. DOI: 10.1145/3170358.3170391.

Mingers, J. and Standing, C. (2020). "A Framework for Validating Information Systems Research Based on a Pluralist Account of Truth and Correctness." In: *Journal of the Association for Information Systems* 21.1, pp. 117–151. DOI: 10.17705/1jais.00594.

Mislevy, R. J. (2016). "How Developments in Psychology and Technology Challenge Validity Argumentation." In: *Journal of Educational Measurement* 53.3, pp. 265–292. DOI: 10.1111/jedm.12117.

Miwa, M., Thomas, J., O'Mara-Eves, A., and Ananiadou, S. (2014). "Reducing Systematic Review Workload through Certainty-Based Screening." In: *Journal of Biomedical Informatics* 51, pp. 242–253. DOI: 10.1016/j.jbi.2014.06.005.

Mohasseb, A., Aziz, B., Jung, J., and Lee, J. (2020). "Cyber Security Incidents Analysis and Classification in a Case Study of Korean Enterprises." In: *Knowledge and Information Systems* 62.7, pp. 2917–2935. DOI: 10.1007/s10115-020-01452-5.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., and PRISMA-P Group (2015). "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement." In: *Systematic Reviews* 4.1, pp. 1–9. DOI: 10.1186/2046-4053-4-1.

Morrison, P., Moye, D., Pandita, R., and Williams, L. (2018). "Mapping the Field of Software Life Cycle Security Metrics." In: *Information and Software Technology* 102, pp. 146–159. DOI: 10.1016/j.infsof.2018.05.011.

Morton, L. W., Eigenbrode, S. D., and Martin, T. A. (2015). "Architectures of Adaptive Integration in Large Collaborative Projects." In: *Ecology and Society* 20.4. JSTOR: 26270306.

Mourão, E., Kalinowski, M., Murta, L., Mendes, E., and Wohlin, C. (2017). "Investigating the Use of a Hybrid Search Strategy for Systematic Reviews." In: *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pp. 193–198. DOI: 10.1109/ESEM.2017.30.

Mourão, E., Pimentel, J. F., Murta, L., Kalinowski, M., Mendes, E., and Wohlin, C. (2020). "On the Performance of Hybrid Search Strategies for Systematic Literature Reviews in Software Engineering." In: *Information and Software Technology* 123, p. 106294. DOI: 10.1016/j.infsof.2020.106294.

Mtsweni, J. S., Shozi, N. A., Matenche, K., Mutemwa, M., Mkhonto, N., and Jansen van Vuuren, J. (2016). "Development of a Semantic-Enabled Cybersecurity Threat Intelligence Sharing Model." In: *11th International Conference on Cyber Warfare & Security*. Boston, MA, USA: Academic Publishing International Ltd.

Muckin, M. and Fitch, S. C. (2019). *A Threat-Driven Approach to Cyber Security*. Tech. rep. Lockheed Martin Corporation, p. 45.

Muñoz, S., Sánchez, E., and Iglesias, C. A. (2020). "An Emotion-Aware Learning Analytics System Based on Semantic Task Automation." In: *Electronics* 9.8, p. 1194. DOI: 10.3390/electronics9081194.

Mutemwa, M., Mtsweni, J., and Mkhonto, N. (2017). "Developing a Cyber Threat Intelligence Sharing Platform for South African Organisations." In: *2017 Conference on Information Communication Technology and Society (ICTAS)*. Durban, South Africa: IEEE, pp. 1–6. DOI: 10.1109/ICTAS.2017.7920657.

NCSC UK (2014). *Cyber Essentials*. https://www.ncsc.gov.uk/cyberessentials/overview.

Nikolopoulos, K. (2021). "We Need to Talk about Intermittent Demand Forecasting." In: *European Journal of Operational Research* 291.2, pp. 549–559. DOI: 10.1016/j.ejor.2019.12.046.

Noel, S. and Jajodia, S. (2014). "Metrics Suite for Network Attack Graph Analytics." In: *Proceedings of the 9th Annual Cyber and Information Security Research Conference*. CISR '14. New York, NY, USA: Association for Computing Machinery, pp. 5–8. DOI: 10.1145/2602087.2602117.

O'Connor, A. M., Tsafnat, G., Thomas, J., Glasziou, P., Gilbert, S. B., and Hutton, B. (2019). "A Question of Trust: Can We Build an Evidence Base to Gain Trust in Systematic Review Automation Technologies?" In: *Systematic Reviews* 8.1, p. 143. DOI: 10.1186/s13643-019-1062-0.

Osborne, F., Muccini, H., Lago, P., and Motta, E. (2019). "Reducing the Effort for Systematic Reviews in Software Engineering." In: *Data Science* 2.1-2, pp. 311–340. DOI: 10.3233/DS-190019.

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). "Rayyan—a Web and Mobile App for Systematic Reviews." In: *Systematic Reviews* 5.1, p. 210. DOI: 10.1186/s13643-016-0384-4.

Padayachee, K. (2012). "Taxonomy of Compliant Information Security Behavior." In: *Computers & Security* 31.5, pp. 673–680. DOI: 10.1016/j.cose.2012.04.004.

Page, M. J. et al. (2021). "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews." In: *Systematic Reviews* 372.1, pp. 1–9. DOI: 10.1186/s13643-021-01626-4.

Pardo, A., Ellis, R. A., and Calvo, R. A. (2015). "Combining Observational and Experiential Data to Inform the Redesign of Learning Activities." In: *Proceedings of the 5th International Learning Analytics and Knowledge Conference*. LAK'15. Poughkeepsie, NY, USA: ACM, pp. 305–309.

Park, Y. and Jo, I.-H. (2019). "Factors That Affect the Success of Learning Analytics Dashboards." In: *Educational Technology Research and Development* 67.6, pp. 1547–1571. DOI: 10.1007/s11423-019-09693-0.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). "A Design Science Research Methodology for Information Systems Research."

In: *Journal of Management Information Systems* 24.3, pp. 45–77. DOI: 10.2753/MIS0742-1222240302.

Pendleton, M., Garcia-Lebron, R., Cho, J.-H., and Xu, S. (2016). "A Survey on Systems Security Metrics." In: *ACM Computing Surveys* 49.4, 62:1–62:35. DOI: 10.1145/3005714.

Petticrew, M. (2001). "Systematic Reviews from Astronomy to Zoology: Myths and Misconceptions." In: *BMJ* 322.7278, pp. 98–101. DOI: 10.1136/bmj.322.7278.98.

Pfleeger, C. P. and Pfleeger, S. L. (2012). *Analyzing Computer Security: A Threat/Vulnerability/Countermeasure Approach*. Prentice Hall Professional.

Pfleeger, S. and Cunningham, R. (2010). "Why Measuring Security Is Hard." In: *IEEE Security & Privacy* 8.4, pp. 46–54. DOI: 10.1109/MSP.2010.60.

Ponemon Institute (2019). *2019 Global State of Cybersecurity in Small and Medium-Sized Businesses*. Technical Report. Keeper Security.

Prat, A. and Code, W. J. (2021). "WeBWorK Log Files as a Rich Source of Data on Student Homework Behaviours." In: *International Journal of Mathematical Education in Science and Technology* 52.10, pp. 1540–1556. DOI: 10.1080/0020739X.2020.1782492.

Pries-Heje, J., Baskerville, R., and Venable, J. (2008). "Strategies for Design Science Research Evaluation." In: *16th European Conference on Information Systems*. Galway, Ireland: Association for Information Systems, p. 87.

Prinsloo, P. and Slade, S. (2015). "Student Privacy Self-Management: Implications for Learning Analytics." In: *Proceedings of the 5th International Learning Analytics & Knowledge Conference*. LAK '15. Poughkeepsie, NY, USA: ACM, pp. 83–92. DOI: 10.1145/2723576.2723585.

Proença, D. and Borbinha, J. (2018). "Information Security Management Systems - A Maturity Model Based on ISO/IEC 27001." In: *Business Information Systems*. Ed. by W. Abramowicz and A. Paschke. Lecture Notes in Business Information Processing. Cham: Springer International Publishing, pp. 102–114. DOI: 10.1007/978-3-319-93931-5_8.

Purohit, S., Calyam, P., Wang, S., Yempalla, R., and Varghese, J. (2020). "DefenseChain: Consortium Blockchain for Cyber Threat Intelligence Sharing and Defense." In: *2020 2nd Conference on Blockchain Research Applications for Innovative Networks and Services (BRAINS)*. Paris, France: IEEE, pp. 112–119. DOI: 10.1109/BRAINS49436.2020.9223313.

Qamar, S., Anwar, Z., Rahman, M. A., Al-Shaer, E., and Chu, B.-T. (2017). "Data-Driven Analytics for Cyber-Threat Intelligence and Information Sharing." In: *Computers & Security* 67, pp. 35–58. DOI: 10.1016/j.cose.2017.02.005.

Radjenović, D., Heričko, M., Torkar, R., and Živkovič, A. (2013). "Software Fault Prediction Metrics: A Systematic Literature Review." In: *Information and Software Technology* 55.8, pp. 1397–1418. DOI: 10.1016/j.infsof.2013.02.009.

Raković, M., Gašević, D., Hassan, S. U., Ruipérez Valiente, J. A., Aljohani, N., and Milligan, S. (2023). "Learning Analytics and Assessment: Emerging Research Trends, Promises and Future Opportunities." In: *British Journal of Educational Technology* 54.1, pp. 10–18. DOI: 10.1111/bjet.13301.

Ramos, A., Lazar, M., Filho, R. H., and Rodrigues, J. J. P. C. (2017). "Model-Based Quantitative Network Security Metrics: A Survey." In: *IEEE Communications Surveys & Tutorials* 19.4, pp. 2704–2734. DOI: 10.1109/COMST.2017.2745505.

Ramsdale, A., Shiaeles, S., and Kolokotronis, N. (2020). "A Comparative Analysis of Cyber-Threat Intelligence Sources, Formats and Languages." In: *Electronics* 9.5, p. 824. DOI: 10.3390/electronics9050824.

Rass, S., Alshawish, A., Abid, M. A., Schauer, S., Zhu, Q., and Meer, H. D. (2017). "Physical Intrusion Games—Optimizing Surveillance by Simulation and Game Theory." In: *IEEE Access* 5, pp. 8394–8407. DOI: 10.1109/ACCESS.2017.2693425.

Ray, J., Marshall, H., De Sousa, V., Jean, J., Warren, S., and Bachand, S. (2020). *2020 Cyber Threatscape Report*. https://www.accenture.com/us-en/insights/security/cyber-threatscape-report.

Recker, J., Zur Muehlen, M., Siau, K., Erickson, J., and Indulska, M. (2009). "Measuring Method Complexity: UML versus BPMN." In: *Proceedings of the 15th Americas Conference on Information Systems*. Ed. by P. Gray, R. Sharda, and R. Nickerson. Online: Association for Information Systems, pp. 1–9.

Refsdal, A., Solhaug, B., and Stølen, K. (2015). *Cyber-Risk Management*. Springer.

Reich, K. (2007). "Interactive Constructivism in Education." In: *Education and Culture* 23.1, pp. 7–26. JSTOR: 42922599.

Riesco, R., Larriva-Novo, X., and Villagra, V. A. (2020). "Cybersecurity Threat Intelligence Knowledge Exchange Based on Blockchain." In: *Telecommunication Systems* 73.2, pp. 259–288. DOI: 10.1007/s11235-019-00613-4.

Riesco, R. and Villagrá, V. A. (2019). "Leveraging Cyber Threat Intelligence for a Dynamic Risk Framework." In: *International Journal of Information Security* 18.6, pp. 715–739. DOI: 10.1007/s10207-019-00433-2.

Ring, T. (2014). "Threat Intelligence: Why People Don't Share." In: *Computer Fraud & Security* 2014.3, pp. 5–9. DOI: 10.1016/S1361-3723(14)70469-5.

Rios, P., Radhakrishnan, A., Williams, C., Ramkissoon, N., Pham, B., Cormack, G. V., Grossman, M. R., Muller, M. P., Straus, S. E., and Tricco, A. C. (2020). "Preventing the Transmission of COVID-19 and Other Coronaviruses in Older Adults Aged 60 Years and above Living in Long-Term Care: A Rapid Review." In: *Systematic Reviews* 9.1, p. 218. DOI: 10.1186/s13643-020-01486-4.

Rittel, H. W. J. and Webber, M. M. (1973). "Dilemmas in a General Theory of Planning." In: *Policy Sciences* 4.2, pp. 155–169. DOI: 10.1007/BF01405730.

Rizvi, S., Rienties, B., and Khoja, S. A. (2019). "The Role of Demographics in Online Learning; A Decision Tree Based Approach." In: *Computers & Education* 137, pp. 32–47. DOI: 10.1016/j.compedu.2019.04.001.

Rodríguez-Triana, M. J., Prieto, L. P., Vozniuk, A., Boroujeni, M. S., Schwendimann, B. A., Holzer, A., and Gillet, D. (2017). "Monitoring, Awareness and Reflection in Blended Technology Enhanced Learning: A Systematic Review." In: *International Journal of Technology Enhanced Learning* 9.2-3, pp. 126–150. DOI: 10.1504/IJTEL.2017.084489.

Ros, R., Bjarnason, E., and Runeson, P. (2017). "A Machine Learning Approach for Semi-Automated Search and Selection in Literature Studies." In: *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. EASE'17. New York, NY, USA: Association for Computing Machinery, pp. 118–127. DOI: 10.1145/3084226.3084243.

Rose, M. E. and Kitchin, J. R. (2019). "Pybliometrics: Scriptable Bibliometrics Using a Python Interface to Scopus." In: *SoftwareX* 10, p. 100263. DOI: 10.1016/j.softx.2019.100263.

Rossiter, E., Thomson, T., and Fitzgerald, R. (2024). "Supporting University Students' Learning across Time and Space: A from-Scratch, Personalised and Mobile-Friendly Approach." In: *Interactive Technology and Smart Education* 21.1, pp. 108–130. DOI: 10.1108/ITSE-07-2022-0082.

Rubel, A. and Jones, K. M. L. (2016). "Student Privacy in Learning Analytics: An Information Ethics Perspective." In: *The Information Society* 32.2, pp. 143–159. DOI: 10.1080/01972243.2016.1130502.

Rudolph, M. and Schwarz, R. (2012). "A Critical Survey of Security Indicator Approaches." In: *2012 Seventh International Conference on Availability, Reliability and Security*, pp. 291–300. DOI: 10.1109/ARES.2012.10.

Ryan, R. M. and Deci, E. L. (2000). "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being." In: *American Psychologist* 55.1, pp. 68–78. DOI: 10.1037/0003-066X.55.1.68.

Sahinoglu, M. (2008). "An Input–Output Measurable Design for the Security Meter Model to Quantify and Manage Software Security Risk." In: *IEEE Transactions on Instrumentation and Measurement* 57.6, pp. 1251–1260. DOI: 10.1109/TIM.2007.915139.

Salehi, R. et al. (2023). "Evaluation of a Continuing Professional Development Strategy on COVID-19 for 10 000 Health Workers in Ghana: A Two-Pronged Approach." In: *Human Resources for Health* 21.1, pp. 1–13. DOI: 10.1186/s12960-023-00804-w.

Saqr, M., Fors, U., and Nouri, J. (2018). "Using Social Network Analysis to Understand Online Problem-Based Learning and Predict Performance." In: *PLOS ONE* 13.9, e0203590. DOI: 10.1371/journal.pone.0203590.

Saqr, M. and López-Pernas, S. (2021). "Modelling Diffusion in Computer-Supported Collaborative Learning: A Large Scale Learning Analytics Study." In: *International Journal of Computer-Supported Collaborative Learning* 16.4, pp. 441–483. DOI: 10.1007/s11412-021-09356-4.

Saqr, M., Viberg, O., and Vartiainen, H. (2020). "Capturing the Participation and Social Dimensions of Computer-Supported Collaborative Learning through Social Network Analysis: Which Method and Measures Matter?"

In: *International Journal of Computer-Supported Collaborative Learning* 15.2, pp. 227–248. DOI: 10.1007/s11412-020-09322-6.

Sarabi, A., Naghizadeh, P., Liu, Y., and Liu, M. (2016). "Risky Business: Fine-grained Data Breach Prediction Using Business Profiles." In: *Journal of Cybersecurity* 2.1, pp. 15–28. DOI: 10.1093/cybsec/tyw004.

Sarker, I. H., Furhad, M. H., and Nowrozy, R. (2021). "AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions." In: *SN Computer Science* 2.3, p. 173. DOI: 10.1007/s42979-021-00557-0.

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., and Ng, A. (2020). "Cybersecurity Data Science: An Overview from Machine Learning Perspective." In: *Journal of Big Data* 7.1, p. 41. DOI: 10.1186/s40537-020-00318-5.

Sauerwein, C., Sillaber, C., Mussmann, A., and Breu, R. (2017). "Threat Intelligence Sharing Platforms: An Exploratory Study of Software Vendors and Research Perspectives." In: *Wirtschaftsinformatik 2017 Proceedings*.

Scandariato, R., Wuyts, K., and Joosen, W. (2015). "A Descriptive Study of Microsoft's Threat Modeling Technique." In: *Requirements Engineering* 20.2, pp. 163–180. DOI: 10.1007/s00766-013-0195-2.

Schaberreiter, T., Kupfersberger, V., Rantos, K., Spyros, A., Papanikolaou, A., Ilioudis, C., and Quirchmayr, G. (2019). "A Quantitative Evaluation of Trust in the Quality of Cyber Threat Intelligence Sources." In: *Proceedings of the 14th International Conference on Availability, Reliability and Security*. ARES 2019. Canterbury, UK: Association for Computing Machinery, pp. 1–10. DOI: 10.1145/3339252.3342112.

Schlette, D., Böhm, F., Caselli, M., and Pernul, G. (2021). "Measuring and Visualizing Cyber Threat Intelligence Quality." In: *International Journal of Information Security* 20.1, pp. 21–38. DOI: 10.1007/s10207-020-00490-y.

Schmidt, S. and Albayrak, S. (2010). "A Quantitative Framework for Dependency-Aware Organizational IT Risk Management." In: *2010 10th International Conference on Intelligent Systems Design and Applications*, pp. 1207–1212. DOI: 10.1109/ISDA.2010.5687022.

Sein, M., Henfridsson, O., Purao, S., Rossi, M., and Lindgren, R. (2011). "Action Design Research." In: *Management Information Systems Quarterly* 35.1, pp. 37–56. DOI: 10.2307/23043488.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). "Fairness and Abstraction in Sociotechnical Systems." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. New York, NY, USA: Association for Computing Machinery, pp. 59–68. DOI: 10.1145/3287560.3287598.

Sengupta, S., Chowdhary, A., Sabur, A., Alshamrani, A., Huang, D., and Kambhampati, S. (2020). "A Survey of Moving Target Defenses for Network Security." In: *IEEE Communications Surveys & Tutorials* 22.3, pp. 1909–1941. DOI: 10.1109/COMST.2020.2982955.

Serketzis, N., Katos, V., Ilioudis, C., Baltatzis, D., and Pangalos, G. J. (2019). "Actionable Threat Intelligence for Digital Forensics Readiness." In: *Information & Computer Security* 27.2, pp. 273–291. DOI: 10.1108/ICS-09-2018-0110.

Settanni, G., Skopik, F., Shovgenya, Y., Fiedler, R., Carolan, M., Conroy, D., Boettinger, K., Gall, M., Brost, G., Ponchel, C., Haustein, M., Kaufmann, H., Theuerkauf, K., and Olli, P. (2017). "A Collaborative Cyber Incident Management System for European Interconnected Critical Infrastructures." In: *Journal of Information Security and Applications* 34, pp. 166–182. DOI: 10.1016/j.jisa.2016.05.005.

Settles, B. (2012). "Active Learning." In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6.1, pp. 1–114. DOI: 10.2200/S00429ED1V01Y201207AIM018.

Settles, B., T. LaFlair, G., and Hagiwara, M. (2020). "Machine Learning–Driven Language Assessment." In: *Transactions of the Association for Computational Linguistics* 8, pp. 247–263. DOI: 10.1162/tacl_a_00310.

Shameli-Sendi, A., Shajari, M., Hassanabadi, M., Jabbarifar, M., and Dagenais, M. (2012). "Fuzzy Multi-Criteria Decision-Making for Information Security Risk Assessment." In: *The Open Cybernetics & Systemics Journal* 6.1.

Shameli-Sendi, A., Aghababaei-Barzegar, R., and Cheriet, M. (2016). "Taxonomy of Information Security Risk Assessment (ISRA)." In: *Computers & Security* 57, pp. 14–30. DOI: 10.1016/j.cose.2015.11.001.

Shemilt, I., Khan, N., Park, S., and Thomas, J. (2016). "Use of Cost-Effectiveness Analysis to Compare the Efficiency of Study Identification Methods in Systematic Reviews." In: *Systematic Reviews* 5.1, p. 140. DOI: 10.1186/s13643-016-0315-4.

Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., and Choi, Y. (2021). "SAINT+: Integrating Temporal Features for EdNet Correctness Prediction." In: *Proceedings of the 11th International Learning Analytics & Knowledge Conference*. LAK '21. Irvine, CA, USA: ACM, pp. 490–496. DOI: 10.1145/3448139.3448188.

Shin, Y., Meneely, A., Williams, L., and Osborne, J. A. (2011). "Evaluating Complexity, Code Churn, and Developer Activity Metrics as Indicators of Software Vulnerabilities." In: *IEEE Transactions on Software Engineering* 37.6, pp. 772–787. DOI: 10.1109/TSE.2010.81.

Shojaifar, A. and Fricker, S. A. (2020). "SMEs' Confidentiality Concerns for Security Information Sharing." In: *Human Aspects of Information Security and Assurance*. Ed. by N. Clarke and S. Furnell. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, pp. 289–299. DOI: 10.1007/978-3-030-57404-8_22.

Shojaifar, A., Fricker, S. A., and Gwerder, M. (2020). "Automating the Communication of Cybersecurity Knowledge: Multi-case Study." In: *Information Security Education. Information Security in Action*. Ed. by L. Drevin, S. Von Solms, and M. Theocharidou. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, pp. 110–124. DOI: 10.1007/978-3-030-59291-2_8.

Shokouhyar, S., Panahifar, F., Karimisefat, A., and Nezafatbakhsh, M. (2018). "An Information System Risk Assessment Model: A Case Study in Online Banking System." In: *International Journal of Electronic Security and Digital Forensics* 10.1, pp. 39–60. DOI: 10.1504/IJESDF.2018.089205.

Siemens, G. (2004). "Connectivism: A Learning Theory for the Digital Age." In: *International Journal of Instructional Technology and Distance Learning* 2.1, pp. 1–8.

Silva, M. M., de Gusmão, A. P. H., Poleto, T., Silva, L. C. e, and Costa, A. P. C. S. (2014). "A Multidimensional Approach to Information Security Risk Management Using FMEA and Fuzzy Theory." In: *International Journal of Information Management* 34.6, pp. 733–740. DOI: 10.1016/j.ijinfomgt.2014.07.005.

Simon, S. (2008). "Using Toulmin's Argument Pattern in the Evaluation of Argumentation in School Science." In: *International Journal of Research & Method in Education* 31.3, pp. 277–289. DOI: 10.1080/17437270802417176.

Singh, P. and Singh, K. (2017). "Exploring Automatic Search in Digital Libraries: A Caution Guide for Systematic Reviewers." In: *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. EASE'17. New York, NY, USA: Association for Computing Machinery, pp. 236–241. DOI: 10.1145/3084226.3084275.

Sinha, T., Jermann, P., Li, N., and Dillenbourg, P. (2014). "Your Click Decides Your Fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: ACL, pp. 3–14. DOI: 10.3115/v1/W14-4102.

SINTEF Digital (2022). *Customer Journey Modeling Language (CJML) v1.1. Licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).* https://www.cjml.no/.

Sittig, D. F. and Singh, H. (2016). "A Socio-Technical Approach to Preventing, Mitigating, and Recovering from Ransomware Attacks." In: *Applied Clinical Informatics* 7.2, pp. 624–632. DOI: 10.4338/ACI-2016-04-SOA-0064.

Skopik, F., Settanni, G., and Fiedler, R. (2016). "A Problem Shared Is a Problem Halved: A Survey on the Dimensions of Collective Cyber Defense through Security Information Sharing." In: *Computers & Security* 60, pp. 154–176. DOI: 10.1016/j.cose.2016.04.003.

Slade, S. and Prinsloo, P. (2013). "Learning Analytics: Ethical Issues and Dilemmas." In: *American Behavioral Scientist* 57.10, pp. 1510–1529. DOI: 10.1177/0002764213479366.

Slayton, R. (2015). "Measuring Risk: Computer Security Metrics, Automation, and Learning." In: *IEEE Annals of the History of Computing* 37.2, pp. 32–45. DOI: 10.1109/MAHC.2015.30.

Society for Learning Analytics Research (SoLAR) (2022). *What Is Learning Analytics?* https://www.solaresearch.org/about/what-is-learning-analytics/.

Spruit, M. and Röling, M. (2014). "ISFAM: The Information Security Focus Area Maturity Model." In: *Proceedings of the European Conference on Information Systems (ECIS) 2014*. Tel Aviv, Israel: AIS.

Stadler, M., Herborn, K., Mustafić, M., and Greiff, S. (2020). "The Assessment of Collaborative Problem Solving in PISA 2015: An Investigation of the Validity of the PISA 2015 CPS Tasks." In: *Computers & Education* 157, pp. 1–11. DOI: 10.1016/j.compedu.2020.103964.

Steinberger, J., Sperotto, A., Golling, M., and Baier, H. (2015). "How to Exchange Security Events? Overview and Evaluation of Formats and Protocols." In: *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*. Ottawa, Canada: IEEE, pp. 261–269. DOI: 10.1109/INM.2015.7140300.

Stergiopoulos, G., Gritzalis, D., and Kouktzoglou, V. (2018). "Using Formal Distributions for Threat Likelihood Estimation in Cloud-Enabled IT Risk Assessment." In: *Computer Networks* 134, pp. 23–45. DOI: 10.1016/j.comnet.2018.01.033.

Stödberg, U. (2012). "A Research Review of E-Assessment." In: *Assessment & Evaluation in Higher Education* 37.5, pp. 591–604. DOI: 10.1080/02602938.2011.557496.

Stolfo, S., Bellovin, S. M., and Evans, D. (2011). "Measuring Security." In: *IEEE Security Privacy* 9.3, pp. 60–65. DOI: 10.1109/MSP.2011.56.

Straub, D. W. (1989). "Validating Instruments in MIS Research." In: *MIS Quarterly* 13.2, pp. 147–169. DOI: 10.2307/248922. JSTOR: 248922.

Straub, D. W., Boudreau, M.-C., and Gefen, D. (2004). "Validation Guidelines for IS Positivist Research." In: *Communications of the Association for Information Systems* 13.1. DOI: 10.17705/1CAIS.01324.

Strauss, A. and Corbin, J. M. (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. 1st. Thousand Oaks, CA, US: SAGE Publications.

Sun, N., Zhang, J., Rimba, P., Gao, S., Zhang, L. Y., and Xiang, Y. (2019). "Data-Driven Cybersecurity Incident Prediction: A Survey." In: *IEEE Communications Surveys & Tutorials* 21.2, pp. 1744–1772. DOI: 10.1109/COMST.2018.2885561.

Sun, Y., Ochiai, H., and Esaki, H. (2020). "Intrusion Detection with Segmented Federated Learning for Large-Scale Multiple LANs." In: *2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow, UK: IEEE, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9207094.

Swiss NCSC (2021). *Cyberthreats*. https://www.ncsc.admin.ch/ncsc/en/home.html.

Tabuenca, B., Kalz, M., Drachsler, H., and Specht, M. (2015). "Time Will Tell: The Role of Mobile Learning Analytics in Self-Regulated Learning." In: *Computers & Education* 89, pp. 53–74. DOI: 10.1016/j.compedu.2015.08.004.

Takahashi, T. and Miyamoto, D. (2016). "Structured Cybersecurity Information Exchange for Streamlining Incident Response Operations." In: *NOMS 2016*

- *2016 IEEE/IFIP Network Operations and Management Symposium*. Istanbul, Turkey: IEEE, pp. 949–954. DOI: 10.1109/NOMS.2016.7502931.

Tanrıverdi, M. and Tekerek, A. (2019). "Implementation of Blockchain Based Distributed Web Attack Detection Application." In: *2019 1st International Informatics and Software Engineering Conference (UBMYK)*. Ankara, Turkey: IEEE, pp. 1–6. DOI: 10.1109/UBMYK48245.2019.8965446.

Thapa, C., Arachchige, P. C. M., Camtepe, S., and Sun, L. (2022). "SplitFed: When Federated Learning Meets Split Learning." In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence*. Vol. 36. Palo Alto, CA, USA: AAAI Press, pp. 8485–8493. DOI: 10.1609/aaai.v36i8.20825.

The VERIS Community Database (2021). Verizon Security Research & Cyber Intelligence Center.

Thornberg, R. (2012). "Informed Grounded Theory." In: *Scandinavian Journal of Educational Research* 56.3, pp. 243–259. DOI: 10.1080/00313831.2011.581686.

Topor, M., Pickering, J., Mendes, A. B., Bishop, D., Büttner, F. C., Elsherif, M., Evans, T. R., Henderson, E. L., Kalandadze, T., and Nitschke, F. (2020). *Non-Interventional, Reproducible, and Open Systematic Review (NIRO-SR) Guidelines*. DOI: 10.17605/OSF.IO/F3BRW.

Toulmin, S. E. (1958). *The Uses of Argument*. 1st ed. Cambridge, UK: Cambridge University Press.

Trend Micro (2021). *Devastating Flubot Malware Spreads From Europe to Australia*.

Tress, G., Tress, B., and Fry, G. (2005). "Clarifying Integrative Research Concepts in Landscape Ecology." In: *Landscape Ecology* 20.4, pp. 479–493. DOI: 10.1007/s10980-004-3290-4.

Tsai, C.-W. (2010). "Do Students Need Teacher's Initiation in Online Collaborative Learning?" In: *Computers & Education* 54.4, pp. 1137–1144. DOI: 10.1016/j.compedu.2009.10.021.

Tucker, B. (2020). "Advancing Risk Management Capability Using the OCTAVE FORTE Process." In: DOI: 10.1184/R1/13014266.v1.

Ural, Ö., Acartürk, C., and Acartürk, C. (2021). "Automatic Detection of Cyber Security Events from Turkish Twitter Stream and Newspaper Data." In: *Proceedings of the 7th International Conference on Information Systems Security and Privacy - ICISSP*. Online, pp. 66–76.

Vakilinia, I., Cheung, S., and Sengupta, S. (2018). "Sharing Susceptible Passwords as Cyber Threat Intelligence Feed." In: *MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM)*. Los Angeles, CA, USA: IEEE, pp. 1–6. DOI: 10.1109/MILCOM.2018.8599742.

Valle, N., Antonenko, P., Valle, D., Sommer, M., Huggins-Manley, A. C., Dawson, K., Kim, D., and Baiser, B. (2021). "Predict or Describe? How Learning Analytics Dashboard Design Influences Motivation and Statistics Anxiety in an Online Statistics Course." In: *Educational Technology Research and Development* 69.3, pp. 1405–1431. DOI: 10.1007/s11423-021-09998-z.

Vallerand, R. J. (1997). "Toward A Hierarchical Model of Intrinsic and Extrinsic Motivation." In: *Advances in Experimental Social Psychology*. Ed. by M. P.

Zanna. Vol. 29. Academic Press, pp. 271–360. DOI: 10.1016/S0065-2601(08)60019-2.

van Aken, J. E. (2013). "Design Science: Valid Knowledge for Socio-technical System Design." In: *Design Science: Perspectives from Europe*. Ed. by M. Helfert and B. Donnellan. Communications in Computer and Information Science. Leixlip, Ireland: Springer, pp. 1–13. DOI: 10.1007/978-3-319-04090-5_1.

van Bavel, R., Rodríguez-Priego, N., Vila, J., and Briggs, P. (2019). "Using Protection Motivation Theory in the Design of Nudges to Improve Online Security Behavior." In: *International Journal of Human-Computer Studies* 123, pp. 29–39. DOI: 10.1016/j.ijhcs.2018.11.003.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., and Oberski, D. L. (2021). "An Open Source Machine Learning Framework for Efficient and Transparent Systematic Reviews." In: *Nature Machine Intelligence* 3.2, pp. 125–133. DOI: 10.1038/s42256-020-00287-7.

van Haastrecht, M., Brinkhuis, M., Peichl, J., Remmele, B., and Spruit, M. (2023). "Embracing Trustworthiness and Authenticity in the Validation of Learning Analytics Systems." In: *Proceedings of the 13th International Learning Analytics and Knowledge Conference*. LAK'23. Arlington, TX, USA: Association for Computing Machinery, pp. 552–558. DOI: 10.1145/3576050.3576060.

van Haastrecht, M., Brinkhuis, M., and Spruit, M. (2024). "Federated Learning Analytics: Investigating the Privacy-Performance Trade-Off in Machine Learning for Educational Analytics." In: *Artificial Intelligence in Education*. Ed. by A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt. Cham: Springer Nature Switzerland, pp. 62–74. DOI: 10.1007/978-3-031-64299-9_5.

van Haastrecht, M., Brinkhuis, M. J. S., and Spruit, M. (2023). *VAST Guideline*. Guideline. https://osf.io/4ygf7/: Leiden University, pp. 1–41.

van Haastrecht, M., Brinkhuis, M. J. S., Wools, S., and Spruit, M. (2023). "VAST: A Practical Validation Framework for e-Assessment Solutions." In: *Information Systems and e-Business Management* 21.1, pp. 603–627. DOI: 10.1007/s10257-023-00641-3.

van Haastrecht, M., Golpur, G., Tzismadia, G., Kab, R., Priboi, C., David, D., Răcătăian, A., Baumgartner, L., Fricker, S., Ruiz, J. F., Armas, E., Brinkhuis, M., and Spruit, M. (2021). "A Shared Cyber Threat Intelligence Solution for SMEs." In: *Electronics* 10.23, p. 2913. DOI: 10.3390/electronics10232913.

van Haastrecht, M., Haas, M., Brinkhuis, M., and Spruit, M. (2024). "Understanding Validity Criteria in Technology-Enhanced Learning: A Systematic Literature Review." In: *Computers & Education* 220, p. 105128. DOI: 10.1016/j.compedu.2024.105128.

van Haastrecht, M., Sarhan, I., Shojaifar, A., Baumgartner, L., Mallouli, W., and Spruit, M. (2021). "A Threat-Based Cybersecurity Risk Assessment Approach Addressing SME Needs." In: *Proceedings of the 16th International*

Conference on Availability, Reliability and Security. ARES 2021. Vienna, Austria: Association for Computing Machinery, pp. 1–12. DOI: 10.1145/3465481.3469199.

van Haastrecht, M., Sarhan, I., Yigit Ozkan, B., Brinkhuis, M., and Spruit, M. (2021). "SYMBALS: A Systematic Review Methodology Blending Active Learning and Snowballing." In: *Frontiers in Research Metrics and Analytics* 6, pp. 1–14. DOI: 10.3389/frma.2021.685591.

van Haastrecht, M., Spruit, M., Schneider, B., Baumgartner, L., Brad, S., Haller, J., Säuberli, R., van Oorschot, T., Remmele, B., and Mallouli, W. (2021). *D4.1 Validation Report*. European Commission Deliverable. GEIGER Consortium, p. 108.

van Haastrecht, M., Yigit Ozkan, B., Brinkhuis, M., and Spruit, M. (2021). "Respite for SMEs: A Systematic Review of Socio-Technical Cybersecurity Metrics." In: *Applied Sciences* 11.15, p. 6909. DOI: 10.3390/app11156909.

Venable, J., Pries-Heje, J., and Baskerville, R. (2016). "FEDS: A Framework for Evaluation in Design Science Research." In: *European Journal of Information Systems* 25.1, pp. 77–89. DOI: 10.1057/ejis.2014.36.

Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., and Duval, E. (2012). "Context-Aware Recommender Systems for Learning: A Survey and Future Challenges." In: *IEEE Transactions on Learning Technologies* 5.4, pp. 318–335. DOI: 10.1109/TLT.2012.11.

Verendel, V. (2009). "Quantified Security Is a Weak Hypothesis: A Critical Survey of Results and Assumptions." In: *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*. NSPW '09. New York, NY, USA: Association for Computing Machinery, pp. 37–50. DOI: 10.1145/1719030.1719036.

Vivekananda-Schmidt, P., Lewis, M., Hassell, A. B., and Group, T. A. V. R. C. R. (2005). "Cluster Randomized Controlled Trial of the Impact of a Computer-Assisted Learning Package on the Learning of Musculoskeletal Examination Skills by Undergraduate Medical Students." In: *Arthritis Care & Research* 53.5, pp. 764–771. DOI: 10.1002/art.21438.

Wagner, C., Dulaunoy, A., Wagener, G., and Iklody, A. (2016). "MISP: The Design and Implementation of a Collaborative Threat Intelligence Sharing Platform." In: *Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security*. WISCS '16. Vienna, Austria: Association for Computing Machinery, pp. 49–56. DOI: 10.1145/2994539.2994542.

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., and Nawaz, R. (2020). "Predicting Academic Performance of Students from VLE Big Data Using Deep Learning Models." In: *Computers in Human Behavior* 104, p. 106189. DOI: 10.1016/j.chb.2019.106189.

Wahono, R. S. (2007). "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks." In: DOI: 10.3923/JSE.2007.1.12.

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., and Trikalinos, T. A. (2012). "Deploying an Interactive Machine Learning System in an Evidence-Based

Practice Center: Abstrackr." In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. IHI '12. New York, NY, USA: Association for Computing Machinery, pp. 819–824. DOI: 10.1145/2110363.2110464.

Wang, C. and Lu, Z. (2018). "Cyber Deception: Overview and the Road Ahead." In: *IEEE Security Privacy* 16.2, pp. 80–85. DOI: 10.1109/MSP.2018.1870866.

Ware, M. and Mabe, M. (2015). *The STM Report: An Overview of Scientific and Scholarly Journal Publishing*. Tech. rep.

Warnat-Herresthal, S. et al. (2021). "Swarm Learning for Decentralized and Confidential Clinical Machine Learning." In: *Nature* 594.7862, pp. 265–270. DOI: 10.1038/s41586-021-03583-3.

Webster, J. and Watson, R. T. (2002). "Analyzing the Past to Prepare for the Future: Writing a Literature Review." In: *MIS Quarterly* 26.2, pp. xiii–xxiii. JSTOR: 4132319.

Whitaker, S., Kinzie, M., Kraft-Sayre, M. E., Mashburn, A., and Pianta, R. C. (2007). "Use and Evaluation of Web-based Professional Development Services Across Participant Levels of Support." In: *Early Childhood Education Journal* 34.6, pp. 379–386. DOI: 10.1007/s10643-006-0142-7.

Whitelock-Wainwright, A., Gašević, D., Tsai, Y.-S., Drachsler, H., Scheffel, M., Muñoz-Merino, P. J., Tammets, K., and Delgado Kloos, C. (2020). "Assessing the Validity of a Learning Analytics Expectation Instrument: A Multinational Study." In: *Journal of Computer Assisted Learning* 36.2, pp. 209–240. DOI: 10.1111/jcal.12401.

Widmer, G. and Kubat, M. (1996). "Learning in the Presence of Concept Drift and Hidden Contexts." In: *Machine Learning* 23.1, pp. 69–101. DOI: 10.1023/A:1018046501280.

Wieringa, R. (2014). *Design science methodology for information systems and software engineering*. Springer. DOI: 10.1007/978-3-662-43839-8.

Wieringa, R. and Moralı, A. (2012). "Technical Action Research as a Validation Method in Information Systems Design Science." In: *Design Science Research in Information Systems. Advances in Theory and Practice*. Ed. by K. Peffers, M. Rothenberger, and B. Kuechler. Lecture Notes in Computer Science. Las Vegas, NV, US: Springer, pp. 220–238. DOI: 10.1007/978-3-642-29863-9_17.

Williamson, B., Bayne, S., and Shay, S. (2020). "The Datafication of Teaching in Higher Education: Critical Issues and Perspectives." In: *Teaching in Higher Education* 25.4, pp. 351–365. DOI: 10.1080/13562517.2020.1748811.

Winne, P. H. (2020). "Construct and Consequential Validity for Learning Analytics Based on Trace Data." In: *Computers in Human Behavior* 112, p. 106457. DOI: 10.1016/j.chb.2020.106457.

Wise, A. F., Vytasek, J. M., Hausknecht, S., and Zhao, Y. (2016). "Developing Learning Analytics Design Knowledge in the "Middle Space": The Student Tuning Model and Align Design Framework for Learning Analytics Use." In: *Online Learning* 20.2, pp. 155–182.

Wohlin, C. (2014). "Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering." In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. EASE '14. New York, NY, USA: Association for Computing Machinery, pp. 1–10. DOI: 10.1145/2601248.2601268.

Wohlin, C., Mendes, E., Felizardo, K. R., and Kalinowski, M. (2020). "Guidelines for the Search Strategy to Update Systematic Literature Reviews in Software Engineering." In: *Information and Software Technology* 127, p. 106366. DOI: 10.1016/j.infsof.2020.106366.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer Science & Business Media.

Wojniusz, S., Thorkildsen, V. D., Heiszter, S. T., and Røe, Y. (2022). "Active Digital Pedagogies as a Substitute for Clinical Placement during the COVID-19 Pandemic: The Case of Physiotherapy Education." In: *BMC Medical Education* 22.1, pp. 1–9. DOI: 10.1186/s12909-022-03916-4.

Woo, Y. and Reeves, T. C. (2007). "Meaningful Interaction in Web-Based Learning: A Social Constructivist Interpretation." In: *The Internet and Higher Education*. Special Section of the AERA Education and World Wide Web Special Interest Group (EdWeb/SIG) 10.1, pp. 15–25. DOI: 10.1016/j.iheduc.2006.10.005.

Wools, S., Eggen, T. J. H. M., and Sanders, P. F. (2010). "Evaluation of Validity and Validation by Means of the Argument-Based Approach." In: *Cadmo* 18.1, pp. 63–82. DOI: 10.3280/CAD2010-001007.

Wools, S., Molenaar, M., and Hopster-den Otter, D. (2019). "The Validity of Technology Enhanced Assessments: Threats and Opportunities." In: *Theoretical and Practical Advances in Computer-based Educational Measurement*. Ed. by B. P. Veldkamp and C. Sluijter. 1st ed. Methodology of Educational Measurement and Assessment. New York, NY, US: Springer, pp. 3–19. DOI: 10.1007/978-3-030-18480-3_1.

Wuyts, K., Scandariato, R., and Joosen, W. (2014). "Empirical Evaluation of a Privacy-Focused Threat Modeling Methodology." In: *Journal of Systems and Software* 96, pp. 122–138. DOI: 10.1016/j.jss.2014.05.075.

Xie, H., Yan, Z., Yao, Z., and Atiquzzaman, M. (2019). "Data Collection for Security Measurement in Wireless Sensor Networks: A Survey." In: *IEEE Internet of Things Journal* 6.2, pp. 2205–2224. DOI: 10.1109/JIOT.2018.2883403.

Xiong, W. and Lagerström, R. (2019). "Threat Modeling – A Systematic Literature Review." In: *Computers & Security* 84, pp. 53–69. DOI: 10.1016/j.cose.2019.03.010.

Yang, J., Wang, Q., Su, C., and Wang, X. (2020). "Threat Intelligence Relationship Extraction Based on Distant Supervision and Reinforcement Learning." In: *32nd International Conference on Software Engineering and Knowledge Engi-*

neering (SEKE 2020). KSIR Virtual Conference Center, USA. DOI: 10.18293/SEKE2020-149.

Yang, N., Singh, T., and Johnston, A. (2020). "A Replication Study of User Motivation in Protecting Information Security Using Protection Motivation Theory and Self Determination Theory." In: *AIS Transactions on Replication Research* 6.1. DOI: 10.17705/1atrr.00053.

Yang, W. and Lam, K.-Y. (2020). "Automated Cyber Threat Intelligence Reports Classification for Early Warning of Cyber Attacks in Next Generation SOC." In: *International Conference on Information and Communications Security (ICICS 2019)*. Ed. by J. Zhou, X. Luo, Q. Shen, and Z. Xu. Lecture Notes in Computer Science. Beijing, China: Springer International Publishing, pp. 145–164. DOI: 10.1007/978-3-030-41579-2_9.

Yang, Y., Shen, J., Qu, Y., Liu, Y., Wang, K., Zhu, Y., Zhang, W., and Yu, Y. (2021). "GIKT: A Graph-Based Interaction Model for Knowledge Tracing." In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by F. Hutter, K. Kersting, J. Lijffijt, and I. Valera. ECML PKDD '20. Ghent, Belgium: Springer International Publishing, pp. 299–315. DOI: 10.1007/978-3-030-67658-2_18.

Ye, D. and Pennisi, S. (2022). "Using Trace Data to Enhance Students' Self-Regulation: A Learning Analytics Perspective." In: *The Internet and Higher Education* 54, p. 100855. DOI: 10.1016/j.iheduc.2022.100855.

Yigit Ozkan, B. and Spruit, M. (2020). "Addressing SME Characteristics for Designing Information Security Maturity Models." In: *Human Aspects of Information Security and Assurance*. Ed. by N. Clarke and S. Furnell. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, pp. 161–174. DOI: 10.1007/978-3-030-57404-8_13.

Yigit Ozkan, B., Spruit, M., Wondolleck, R., and Burriel Coll, V. (2019). "Modelling Adaptive Information Security for SMEs in a Cluster." In: *Journal of Intellectual Capital* 21.2, pp. 235–256. DOI: 10.1108/JIC-05-2019-0128.

Yigit Ozkan, B., van Lingen, S., and Spruit, M. (2021). "The Cybersecurity Focus Area Maturity (CYSFAM) Model." In: *Journal of Cybersecurity and Privacy* 1.1, pp. 119–139. DOI: 10.3390/jcp1010007.

Yoo, M. and Jin, S.-H. (2020). "Development and Evaluation of Learning Analytics Dashboards to Support Online Discussion Activities." In: *Educational Technology & Society* 23.2, pp. 1–18.

You, Y., Oh, S., and Lee, K. (2015). "Advanced Security Assessment for Control Effectiveness." In: *Information Security Applications*. Ed. by K.-H. Rhee and J. H. Yi. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 383–393. DOI: 10.1007/978-3-319-15087-1_30.

Yu, Z., Barik, T., and Menzies, T. (2018). *Fastread/Src: FAST2 Data Update*.

– (2020). *Fastread/Src: Latest Labeled*. Zenodo.

Yu, Z., Kraft, N. A., and Menzies, T. (2018). "Finding Better Active Learners for Faster Literature Reviews." In: *Empirical Software Engineering* 23.6, pp. 3161–3186. DOI: 10.1007/s10664-017-9587-0.

Yu, Z. and Menzies, T. (2019). "FAST2: An Intelligent Assistant for Finding Relevant Papers." In: *Expert Systems with Applications* 120, pp. 57–71. DOI: 10.1016/j.eswa.2018.11.021.

Zhai, X., Krajcik, J., and Pellegrino, J. W. (2021). "On the Validity of Machine Learning-based Next Generation Science Assessments: A Validity Inferential Network." In: *Journal of Science Education and Technology* 30.2, pp. 298–312. DOI: 10.1007/s10956-020-09879-9.

Zhang, H. and Ali Babar, M. (2013). "Systematic Reviews in Software Engineering: An Empirical Investigation." In: *Information and Software Technology* 55.7, pp. 1341–1354. DOI: 10.1016/j.infsof.2012.09.008.

Zhang, H., Babar, M. A., and Tell, P. (2011). "Identifying Relevant Studies in Software Engineering." In: *Information and Software Technology*. Special Section: Best Papers from the APSEC 53.6, pp. 625–637. DOI: 10.1016/j.infsof.2010.12.010.

Zhao, H. and Silverajan, B. (2020). "A Dynamic Visualization Platform for Operational Maritime Cybersecurity." In: *Cooperative Design, Visualization, and Engineering*. Ed. by Y. Luo. Lecture Notes in Computer Science. Bangkok, Thailand: Springer International Publishing, pp. 202–208. DOI: 10.1007/978-3-030-60816-3_23.

Zhao, J., Yan, Q., Li, J., Shao, M., He, Z., and Li, B. (2020). "TIMiner: Automatically Extracting and Analyzing Categorized Cyber Threat Intelligence from Social Data." In: *Computers & Security* 95, p. 101867. DOI: 10.1016/j.cose.2020.101867.

Zhao, Y., Lang, B., and Liu, M. (2017). "Ontology-Based Unified Model for Heterogeneous Threat Intelligence Integration and Sharing." In: *2017 11th IEEE International Conference on Anti-counterfeiting, Security, and Identification (ASID)*. Xiamen, China: IEEE, pp. 11–15. DOI: 10.1109/ICASID.2017.8285734.

Zheng, L., Niu, J., and Zhong, L. (2022). "Effects of a Learning Analytics-Based Real-Time Feedback Approach on Knowledge Elaboration, Knowledge Convergence, Interactive Relationships and Group Performance in CSCL." In: *British Journal of Educational Technology* 53.1, pp. 130–149.

Zhou, X., Jin, Y., Zhang, H., Li, S., and Huang, X. (2016). "A Map of Threats to Validity of Systematic Literature Reviews in Software Engineering." In: *2016 23rd Asia-Pacific Software Engineering Conference (APSEC)*, pp. 153–160. DOI: 10.1109/APSEC.2016.031.

Zhou, Y., Zhang, H., Huang, X., Yang, S., Babar, M. A., and Tang, H. (2015). "Quality Assessment of Systematic Reviews in Software Engineering: A Tertiary Study." In: *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*. EASE '15. New York, NY, USA:

Association for Computing Machinery, pp. 1–14. DOI: 10.1145/2745802.2745815.

Zibak, A. and Simpson, A. (2019). "Cyber Threat Information Sharing: Perceived Benefits and Barriers." In: *Proceedings of the 14th International Conference on Availability, Reliability and Security*. ARES '19. New York, NY, USA: Association for Computing Machinery, pp. 1–9. DOI: 10.1145/3339252.3340528.

Zimmermann, V. and Renaud, K. (2019). "Moving from a 'Human-as-Problem' to a 'Human-as-Solution' Cybersecurity Mindset." In: *International Journal of Human-Computer Studies*. 50 Years of the International Journal of Human-Computer Studies. Reflections on the Past, Present and Future of Human-Centred Technologies 131, pp. 169–187. DOI: 10.1016/j.ijhcs.2019.05.005.

Zumbo, B. D., Maddox, B., and Care, N. M. (2023). "Process and Product in Computer-Based Assessments: Clearing the Ground for a Holistic Validity Framework." In: *European Journal of Psychological Assessment* 39.4, pp. 252–262. DOI: 10.1027/1015-5759/a000748.

LIST OF PUBLICATIONS

Includes first-author deliverables of the GEIGER Horizon 2020 project.

2021

de Vicente Mohino, J. J., Mallouli, W., Ruiz, J. F., and van Haastrecht, M. (2021). "GEIGER: Solution for Small Businesses to Protect Themselves against Cyber-Threats." In: *Proceedings of the 16th International Conference on Availability, Reliability and Security*. ARES '21. Vienna, Austria: Association for Computing Machinery, pp. 1–4. DOI: 10.1145/3465481.3469202.

Smit, T., van Haastrecht, M., and Spruit, M. (2021). "The Effect of Countermeasure Readability on Security Intentions." In: *Journal of Cybersecurity and Privacy* 1.4, pp. 675–703. DOI: 10.3390/jcp1040034.

van Haastrecht, M., Golpur, G., Tzismadia, G., Kab, R., Priboi, C., David, D., Răcătăian, A., Baumgartner, L., Fricker, S., Ruiz, J. F., Armas, E., Brinkhuis, M., and Spruit, M. (2021). "A Shared Cyber Threat Intelligence Solution for SMEs." In: *Electronics* 10.23, p. 2913. DOI: 10.3390/electronics10232913.

van Haastrecht, M., Sarhan, I., Shojaifar, A., Baumgartner, L., Mallouli, W., and Spruit, M. (2021). "A Threat-Based Cybersecurity Risk Assessment Approach Addressing SME Needs." In: *Proceedings of the 16th International Conference on Availability, Reliability and Security*. ARES 2021. Vienna, Austria: Association for Computing Machinery, pp. 1–12. DOI: 10.1145/3465481.3469199.

van Haastrecht, M., Sarhan, I., Yigit Ozkan, B., Brinkhuis, M., and Spruit, M. (2021). "SYMBALS: A Systematic Review Methodology Blending Active Learning and Snowballing." In: *Frontiers in Research Metrics and Analytics* 6, pp. 1–14. DOI: 10.3389/frma.2021.685591.

van Haastrecht, M., Spruit, M., Schneider, B., Baumgartner, L., Brad, S., Haller, J., Säuberli, R., van Oorschot, T., Remmele, B., and Mallouli, W. (2021). *D4.1 Validation Report*. European Commission Deliverable. GEIGER Consortium, p. 108.

van Haastrecht, M., Yigit Ozkan, B., Brinkhuis, M., and Spruit, M. (2021). "Respite for SMEs: A Systematic Review of Socio-Technical Cybersecurity Metrics." In: *Applied Sciences* 11.15, p. 6909. DOI: 10.3390/app11156909.

2022

van Haastrecht, M., Spruit, M., Fricker, S., Schneider, B., Jonkers, N., Löffler, E., Kern, D., Haller, J., Säuberli, R., van Oorschot, T., Kuper, J., Brad, S., Oprisa, C., Campian, D., Remmele, B., Järvinen, H., Mallouli, W., Nguyen, L., and Armas, E. (2022). *D4.2 Demonstration Report*. European Commission Deliverable. GEIGER Consortium, p. 53.

2023

Ferguson, R. et al. (2023). "Aligning the Goals of Learning Analytics with Its Research Scholarship: An Open Peer Commentary Approach." In: *Journal of Learning Analytics* 10.2, pp. 14–50. DOI: 10.18608/jla.2023.8197.

van Haastrecht, M., Brinkhuis, M., Peichl, J., Remmele, B., and Spruit, M. (2023). "Embracing Trustworthiness and Authenticity in the Validation of Learning Analytics Systems." In: *Proceedings of the 13th International Learning Analytics and Knowledge Conference*. LAK'23. Arlington, TX, USA: Association for Computing Machinery, pp. 552–558. DOI: 10.1145/3576050.3576060.

van Haastrecht, M., Brinkhuis, M. J. S., and Spruit, M. (2023). *VAST Guideline*. Guideline. https://osf.io/4ygf7/: Leiden University, pp. 1–41.

van Haastrecht, M., Brinkhuis, M. J. S., Wools, S., and Spruit, M. (2023). "VAST: A Practical Validation Framework for e-Assessment Solutions." In: *Information Systems and e-Business Management* 21.1, pp. 603–627. DOI: 10.1007/s10257-023-00641-3.

2024

van Haastrecht, M., Brinkhuis, M., and Spruit, M. (2024). "Federated Learning Analytics: Investigating the Privacy-Performance Trade-Off in Machine Learning for Educational Analytics." In: *Artificial Intelligence in Education*. Ed. by A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt. Cham: Springer Nature Switzerland, pp. 62–74. DOI: 10.1007/978-3-031-64299-9_5.

van Haastrecht, M., Haas, M., Brinkhuis, M., and Spruit, M. (2024). "Understanding Validity Criteria in Technology-Enhanced Learning: A Systematic Literature Review." In: *Computers & Education* 220, p. 105128. DOI: 10.1016/j.compedu.2024.105128.

SUMMARY IN ENGLISH

Technologies that help to enhance our educational environments can be found everywhere. Some examples are the electronic whiteboards and tablets that enhance classroom interaction, online peer feedback platforms that enhance student collaboration, and, as was the case in the GEIGER project forming the foundation for this dissertation, a mobile application that enhances the educational experience of SMEs trying to learn about cybersecurity. A question we can ask about technology-enhanced learning (TEL) environments is: Do they achieve what they were intended to achieve? We tackle this question by collecting, analysing, and structuring insights through a transdisciplinary approach resulting in artefacts with the potential to impact science and society.

When we ask what a TEL solution is intended to achieve, we touch on the phases of problem investigation and treatment design that form the first two elements of the engineering cycle. As part of our problem investigation phase, we develop a systematic review methodology called SYMBALS (Chapter 2), which we then use to investigate the problem domain for the TEL use case of GEIGER (Chapter 3). We use our findings as input for the treatment design phase, where we employ insights from behavioural theory to design an educational cybersecurity application for SMEs (Chapter 4) and demonstrate experimentally how this application can use external cyber threat intelligence to enhance the educational experience of users (Chapter 5).

In the third phase of the engineering cycle – treatment validation - we turn to the question: How exactly can we show that an intervention achieves what it was intended to achieve? In technical terms, how can we argue for the validity of a TEL intervention. Validity is a multi-faceted concept which is treated differently in different academic disciplines, and we need to recognise it as such. We begin by building a case for taking a holistic perspective in the validation of TEL, supported by a review of the literature and an epistemological analysis (Chapter 6). We expand on this case by conducting a systematic review to improve our understanding of the validity criteria landscape in TEL (Chapter 7), and then combine our earlier insights with a multi-grounded action research approach to develop a comprehensive validation framework for TEL solutions (Chapter 8).

Treatment implementation constitutes the fourth and final phase of the engineering cycle; a phase that can potentially initiate a new cycle, with a new problem and new research questions. We demonstrate through technical experiments and expert interviews how federated learning, a privacy-preserving machine learning technique, could yield an improved implementation of solutions such as GEIGER, by preserving student privacy in educational environments (Chapter 9).

This dissertation represents a pivotal first step towards holistic TEL validation. Validation that aids accelerated, but also responsible and trustworthy, impact. If our validation strategies are misguided, our innovations will follow this misguided path. We cannot accept such a future.

SAMENVATTING IN HET NEDERLANDS

Technologieën die onze onderwijsomgevingen verbeteren zijn overal te vinden. Enkele voorbeelden zijn elektronische whiteboards en tablets die interactie in de klas verbeteren, online peer feedback platforms die de samenwerking tussen studenten vergemakkelijken, en, zoals het geval was in het GEIGER project dat de basis vormt voor dit proefschrift, een mobiele applicatie die de educatieve ervaring verbetert van MKBers die meer willen leren over cybersecurity. Een vraag die men kan stellen over technologisch-ondersteunde leeromgevingen (TEL) is: Bereiken wij hiermee de doelen die wij wilden bereiken? Wij beantwoorden deze vraag door inzichten te verzamelen, te analyseren en te structureren via een transdisciplinaire aanpak, resulterend in artefacten met de potentie om zowel op de wetenschap als de maatschappij impact te hebben.

Wanneer we vragen wat wij met een TEL oplossing willen bereiken, raken we aan de fases van probleem-inventarisatie en ontwerp. Dit zijn de eerste twee fases van de engineering-cyclus. Als onderdeel van de probleem-inventarisatie ontwikkelen wij in dit proefschrift een systematische review methodologie genaamd SYMBALS (Hoofdstuk 2), die we vervolgens gebruiken om het probleemdomein te onderzoeken voor de TEL casus van GEIGER (Hoofdstuk 3). We gebruiken onze bevindingen als inbreng voor Deel ii, dat gaat over het ontwerp van GEIGER. We gebruiken inzichten uit de gedragstheorie om een educatieve cybersecurity-applicatie voor het MKB te ontwerpen (Hoofdstuk 4), en tonen experimenteel aan hoe deze applicatie externe informatie over cyberdreigingen kan gebruiken om de educatieve ervaring van gebruikers te verbeteren (Hoofdstuk 5).

In de derde fase van de engineering-cyclus, de validatie fase, richten we ons op de vraag: Hoe kunnen we precies aantonen dat een interventie het doel bereikt wat het had moeten bereiken? In vaktermen: Hoe kunnen we de validiteit van een TEL interventie beargumenteren? Validiteit is een veelzijdig concept dat verschillend wordt behandeld in verschillende academische disciplines, en we moeten het ook als zodanig erkennen. In Hoofdstuk 6 beargumenteren wij de noodzaak voor een holistisch perspectief bij de validatie van TEL, ondersteund door een overzicht van de literatuur en een epistemologische analyse. We breiden dit argument uit middels een systematische review die ons inzicht geeft in het landschap van validiteitscriteria voor TEL (Hoofdstuk 7). Vervolgens combineren we deze inzichten met een multi-grounded action research benadering om een validatieraamwerk voor TEL oplossingen te ontwikkelen (Hoofdstuk 8).

De implementatie vormt de vierde en laatste fase van de engineering-cyclus; een fase die mogelijk een nieuwe cyclus kan initiëren, met een nieuwe uitdag-

ing en nieuwe onderzoeksvragen. Wij laten door middel van experimenten en interviews met experts zien hoe federated learning, een privacy-beschermende machine learning techniek, een verbeterde implementatie van oplossingen zoals GEIGER zou kunnen opleveren, door de privacy van studenten te beschermen (Hoofdstuk 9).

Dit proefschrift vormt een cruciale eerste stap in de richting van holistische TEL-validatie. Validatie die versnelde, verantwoorde en betrouwbare impact faciliteert. Als onze validatiestrategieën de verkeerde prikkels bevatten, zullen onze innovaties dit verkeerde pad volgen. Een dergelijke toekomst kunnen wij ons niet veroorloven.

CURRICULUM VITAE

Max van Haastrecht was born in Bloemendaal, The Netherlands on the 9th of October 1995. He lived in Germany and Ireland during his youth, before returning to The Netherlands and commencing his university study. He completed the BSc Econometrics and Operations Research (2013-2017, Cum Laude) and MSc Econometrics, Operations Research, and Actuarial Studies (2017-2018, Cum Laude) at the Rijksuniversiteit Groningen. For his MSc thesis, he spent eight months working as an intern at Ortec B.V. in Zoetermeer. He worked as a fraud detection data analyst at Rabobank for 1.5 years, before starting his PhD under the supervision of promotor prof.dr. Marco Spruit and co-promotor dr. Matthieu Brinkhuis at Utrecht University in June of 2020. He followed his promotor to Leiden University in October of 2021, where he completed his PhD and was a member of the University Council on behalf of PhDoc (2022-2024).

SIKS DISSERTATION SERIES

2016	01	Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
	02	Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
	03	Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
	04	Laurens Rietveld (VUA), Publishing and Consuming Linked Data
	05	Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
	06	Michel Wilson (TUD), Robust scheduling in an uncertain environment
	07	Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
	08	Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
	09	Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
	10	George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
	11	Anne Schuth (UvA), Search Engines that Learn from Their Users
	12	Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
	13	Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
	14	Ravi Khadka (UU), Revisiting Legacy Software System Modernization
	15	Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
	16	Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
	17	Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
	18	Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
	19	Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
	20	Daan Odijk (UvA), Context & Semantics in News & Web Search
	21	Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
	22	Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
	23	Fei Cai (UvA), Query Auto Completion in Information Retrieval
	24	Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
	25	Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
	26	Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
	27	Wen Li (TUD), Understanding Geo-spatial Information on Social Media
	28	Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
	29	Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
	30	Ruud Mattheij (TiU), The Eyes Have It
	31	Mohammad Khelghati (UT), Deep web content monitoring
	32	Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
	33	Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
	34	Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
	35	Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
	36	Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
	37	Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
	38	Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
	39	Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
	40	Christian Detweiler (TUD), Accounting for Values in Design
	41	Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
	42	Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
	43	Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
	44	Thibault Sellam (UvA), Automatic Assistants for Database Exploration
	45	Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
	46	Jorge Gallego Perez (UT), Robots to Make you Happy

	47	Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
	48	Tanja Buttler (TUD), Collecting Lessons Learned
	49	Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
	50	Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains
2017	01	Jan-Jaap Oerlemans (UL), Investigating Cybercrime
	02	Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
	03	Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
	04	Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
	05	Mahdieh Shadi (UvA), Collaboration Behavior
	06	Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
	07	Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
	08	Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
	09	Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
	10	Robby van Delden (UT), (Steering) Interactive Play Behavior
	11	Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
	12	Sander Leemans (TU/e), Robust Process Mining with Guarantees
	13	Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
	14	Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
	15	Peter Berck (RUN), Memory-Based Text Correction
	16	Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
	17	Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
	18	Ridho Reinanda (UvA), Entity Associations for Search
	19	Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
	20	Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
	21	Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
	22	Sara Magliacane (VUA), Logics for causal inference under uncertainty
	23	David Graus (UvA), Entities of Interest — Discovery in Digital Traces
	24	Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
	25	Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
	26	Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
	27	Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
	28	John Klein (VUA), Architecture Practices for Complex Contexts
	29	Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
	30	Wilma Latuny (TiU), The Power of Facial Expressions
	31	Ben Ruijl (UL), Advances in computational methods for QFT calculations
	32	Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
	33	Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
	34	Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
	35	Martine de Vos (VUA), Interpreting natural science spreadsheets
	36	Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
	37	Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
	38	Alex Kayal (TUD), Normative Social Applications
	39	Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
	40	Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
	41	Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
	42	Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
	43	Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
	44	Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
	45	Bas Testerink (UU), Decentralized Runtime Norm Enforcement
	46	Jan Schneider (OU), Sensor-based Learning Support

	47	Jie Yang (TUD), Crowd Knowledge Creation Acceleration
	48	Angel Suarez (OU), Collaborative inquiry-based learning
2018	01	Han van der Aa (VUA), Comparing and Aligning Process Representations
	02	Felix Mannhardt (TU/e), Multi-perspective Process Mining
	03	Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
	04	Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
	05	Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
	06	Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
	07	Jieting Luo (UU), A formal account of opportunism in multi-agent systems
	08	Rick Smetsers (RUN), Advances in Model Learning for Software Systems
	09	Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
	10	Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
	11	Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
	12	Xixi Lu (TU/e), Using behavioral context in process mining
	13	Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
	14	Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
	15	Naser Davarzani (UM), Biomarker discovery in heart failure
	16	Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
	17	Jianpeng Zhang (TU/e), On Graph Sample Clustering
	18	Henriette Nakad (UL), De Notaris en Private Rechtspraak
	19	Minh Duc Pham (VUA), Emergent relational schemas for RDF
	20	Manxia Liu (RUN), Time and Bayesian Networks
	21	Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
	22	Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
	23	Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
	24	Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
	25	Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
	26	Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
	27	Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
	28	Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
	29	Yu Gu (TiU), Emotion Recognition from Mandarin Speech
	30	Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
2019	01	Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
	02	Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
	03	Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
	04	Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
	05	Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
	06	Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
	07	Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
	08	Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
	09	Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
	10	Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
	11	Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
	12	Jacqueline Heinerman (VUA), Better Together
	13	Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
	14	Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
	15	Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
	16	Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
	17	Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
	18	Gerard Wagenaar (UU), Artefacts in Agile Team Communication
	19	Vincent Koeman (TUD), Tools for Developing Cognitive Agents
	20	Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
	21	Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
	22	Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture

	23	Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
	24	Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
	25	Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
	26	Prince Singh (UT), An Integration Platform for Synchromodal Transport
	27	Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
	28	Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
	29	Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
	30	Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
	31	Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
	32	Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
	33	Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
	34	Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
	35	Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
	36	Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
	37	Jian Fang (TUD), Database Acceleration on FPGAs
	38	Akos Kadar (OU), Learning visually grounded and multilingual representations
2020	01	Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
	02	Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
	03	Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
	04	Maarten van Gompel (RUN), Context as Linguistic Bridges
	05	Yulong Pei (TU/e), On local and global structure mining
	06	Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
	07	Wim van der Vegt (OU), Towards a software architecture for reusable game components
	08	Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
	09	Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
	10	Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
	11	Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
	12	Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
	13	Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
	14	Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
	15	Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
	16	Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
	17	Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
	18	Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
	19	Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
	20	Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
	21	Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
	22	Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
	23	Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
	24	Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
	25	Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
	26	Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
	27	Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
	28	Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
	29	Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
	30	Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
	31	Gongjin Lan (VUA), Learning better – From Baby to Better
	32	Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
	33	Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
	34	Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
	35	Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021	01	Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
	02	Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
	03	Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
	04	Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
	05	Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
	06	Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
	07	Armel Lefebvre (UU), Research data management for open science
	08	Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
	09	Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
	10	Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
	11	Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
	12	Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
	13	Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
	14	Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
	15	Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
	16	Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
	17	Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
	18	Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
	19	Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
	20	Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
	21	Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
	22	Sihang Qiu (TUD), Conversational Crowdsourcing
	23	Hugo Manuel Proença (UL), Robust rules for prediction and description
	24	Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
	25	Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
	26	Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
	27	Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
	28	Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
2022	01	Judith van Stegeren (UT), Flavor text generation for role-playing video games
	02	Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
	03	Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
	04	Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
	05	Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
	06	Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
	07	Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
	08	Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
	09	Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
	10	Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
	11	Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
	12	Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
	13	Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
	14	Michiel Overeem (UU), Evolution of Low-Code Platforms
	15	Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
	16	Pieter Gijsbers (TU/e), Systems for AutoML Research
	17	Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
	18	Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
	19	Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
	20	Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
	21	Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
	22	Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
	23	Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents

	24	Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
	25	Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
	26	Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
	27	Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
	28	Onuralp Ulusoy (UU), Privacy in Collaborative Systems
	29	Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
	30	Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
	31	Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
	32	Cezara Pastrav (UU), Social simulation for socio-ecological systems
	33	Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
	34	Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
	35	Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction
2023	01	Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
	02	Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
	03	Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
	04	Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
	05	Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
	06	António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
	07	Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
	08	Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
	09	Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
	10	Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
	11	Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
	12	Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
	13	Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
	14	Selma Čaušević (TUD), Energy resilience through self-organization
	15	Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
	16	Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
	17	Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
	18	Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
	19	George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
	20	Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
	21	Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
	22	Alireza Shojaifar (UU), Volitional Cybersecurity
	23	Theo Theunissen (UU), Documentation in Continuous Software Development
	24	Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
	25	Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
	26	Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
	27	Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
	28	Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
	29	Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results
2024	01	Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
	02	Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
	03	Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
	04	Mike Huisman (UL), Understanding Deep Meta-Learning
	05	Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
	06	Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
	07	Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
	08	Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation

09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks

10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science

11 withdrawn

12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning

13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence

14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling

15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health

16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems

17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design

18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations

19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior

20 Ritsart Anne Plantenga (UL), Omgang met Regels

21 Federica Vinella (UU), Crowdsourcing User-Centered Teams

22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts

23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution

24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour

25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics

26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework

27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning

28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs

29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases

30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting

31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows

32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining

33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method

35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach

36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs

37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models

38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings

39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance

40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models

41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines

42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis

43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms

44 Marcel Schmitz (OU), Towards learning analytics-supported learning design

45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making

46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology