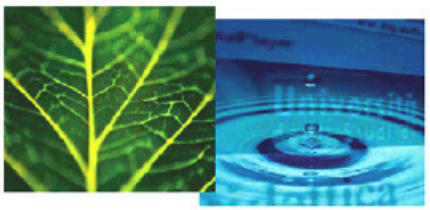
PhD Dissertation



International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Security Risk Assessment Methods: An Evaluation Framework and Theoretical Model of the Criteria Behind Methods' Success

Katsiaryna Labunets

Advisor:

Prof. Fabio Massacci

Università degli Studi di Trento

April 2016

Acknowledgements

Doing my Ph.D. and writing this thesis became a great challenge for me. It would not have been possible to walk this path without support of my family, colleagues and friends.

I am thankful to Prof. Fabio Massacci for giving me the opportunity to work under his supervision and guiding me on my way in the world of science. I owe special thanks to my co-supervisor and main co-author Prof. Federica Paci for her endless and wise support. I would also like to thank the professors who served on my thesis committee: Prof. Maya Daneva (University of Twente), Prof. Paolo Giorgini (University of Trento), Prof. Narayan Ramasubbu (University of Pittsburgh), and Dr. Paolo Tonella (FBK).

Finally, my deepest thanks to my mother Ludmila for her silent patience; my sister Darya for her optimism; my dear Sasha for his unconditional love, greatest support and understanding all these years; and my cat Romashka for keeping my life under her "purrfect" control.

Abstract

Over the past decades a significant number of methods to identify and mitigate security risks have been proposed, but there are few empirical evaluations that show whether these methods are actually effective. So how can practitioners decide which method is the best for security risk assessment of their projects?

To this end, we propose an evaluation framework to compare security risk assessment methods that evaluates the quality of results of methods application with help of external industrial experts and can identify aspects having an effect on the successful application of these methods. The results of the framework application helped us to build the model of key aspects that impact the success of a security risk assessment. Among these aspects are i) the use of catalogues of threats and security controls which can impact methods' actual effectiveness and perceived usefulness and ii) the use of visual representation of risk models that can positively impact methods' perceived ease of use, but negatively affect methods' perceived usefulness if the visual representation is not comprehensible due to scalability issues. To further investigate these findings, we conducted additional empirical investigations: i) how different features of the catalogues of threats and security controls contribute into an effective risk assessment process for novices and experts in either domain or security knowledge, and ii) how comprehensible are different representation approaches for risk models (e.g. tabular and graphical).

Keywords
Security risk assessment; empirical comparison; controlled experiments; security catalogues; risk model comprehensibility.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

An increasing role of security in software development process is recognized by both industrial professionals [103] and academia members [73]. Security risk assessment (SRA) plays an essential role in delivering of secure software systems and should be used from the very beginning of the software development process to eliminate the costly redesign of the system due to emerging issues.

In 2015 PricewaterhouseCoopers' survey[1] reported that 91% of the companies had adopted a risk-based security method (e.g. ISO 27001 or National Institute of Standards and Technology (NIST) Cybersecurity Framework). However, the average total financial loss due to security incidents was decreased only by 5% (from $2.7 in 2014 down to $2.5m in 2015) while information security budgets were increased by 24%. It is still doubtful if current security methods work and worth to adopt, and we need to understand *which security methods are actually effective?*

Recently, a lot of security methods have been proposed by both industry (UK IA's, NIST SP 800-30, BSI 100-1, ISO 2700x, SABSA, etc.) and academia (CORAS, SREP, LINDDUN, Secure Tropos, etc.) Most of them share a similar process: 1) identifying assets, 2) identifying threats to the assets, 3) evaluating risk level of the threats, and 4) identifying security controls to mitigate the risks. The methods may look effective and easy to use on the paper, but the real experience can be different. *How can we validate and compare different SRA methods?* To this end, we need to find an appropriate framework.

A number of theories has been proposed aiming to explain the factors affecting the acceptance of a technology by users: Technology Acceptance Model (TAM) [14], Motivational Model [15], Innovation Diffusion Theory [49], and others. Later, [124] combined the models mentioned above and proposed the Unified Theory of Acceptance and Use of Technology. At the same time Moody proposed Method Evaluation Model (MEM) [77] based on TAM. We found MEM more suitable for our research purposes because it pro-

[1] PricewaterhouseCoopers The Global State of Information Security: Survey 2016. `http://www.pwc.com/gx/en/issues/cyber-security/information-security-survey.html`. Last accessed in March 2016.

vides a theoretical model and measurable constructs for evaluating IS methods. However, due to the generality, the model is missing the factors that are specific to SRA nature and cannot explain its success. Thus, another question arises: *what are the criteria that define the success of an SRA method?*

1.1 Research Questions

Based on the discussion above, this thesis focuses on the following research questions:

RQ1. How can we validate and compare different SRA methods? In the literature there are three main research streams related to the comparison of methods and standards for identifying threats and security controls. The first research stream is represented by those works where methodologies for identifying threats and security controls are proven to be successful by simply showing that they can be applied to a more or less realistic scenario [71,121]. The second stream is related to comparative reviews of methods based on theoretical frameworks that specify criterion for comparison [13, 22, 85, 126]. The third stream is related to works were SRA methods are evaluated through controlled experiments or case studies where MSc students and/or practitioners apply the methods and then provide feedbacks on how these methods are successful in obtaining their goals [20, 51, 64, 65, 86, 123]. However, *there is no well-defined framework* for comparison and validation of SRA methods.

RQ2. Which SRA methods are actually effective? Very few methods and techniques proposed in academia are validated in empirical studies. According to the survey conducted in 2009 [8], only 13% of research works in Requirements Engineering relied on case studies. Recently Cruz et al. [11] revealed that 38.95% of the papers in Cloud Computing Security area propose solutions that have not been evaluated in real practical scenario and only 3.78% of the papers report the evaluation research where the proposed solution has been implemented in practical settings (e.g. case study). Disregarding the validation activities is a double-edged sword: i) practitioners do not know which methods to apply in projects because designers of methods do not provide information about methods' effectiveness and usefulness in real cases and ii) methods designers do not know whether their methods are efficient in practice or not because there is no experience in practical application of the methods. For this reason a series of empirical studies comparing and validating different SRA methods should be conducted.

RQ3. What criteria define success of an SRA method? In order to help security practitioners in selection and method designers in improvement of SRA methods, it is important to have a clear model explaining what are the aspects affecting success of an SRA method and how.

1.2 Thesis Contributions

This section briefs the major contributions of the dissertation.

- **An evaluation framework** that defines a formal procedure for comparison of SRA methods. In particular, the framework evaluates the quality of results of methods application with help of external industry experts and can identify aspects affecting the successful application of these methods.
- **An empirical comparison of different SRA methods** using proposed evaluation framework. The conducted experiments assessed 6 different methods. The experimental results revealed that threat-based methods were perceived by the participants to be superior to other goal- and problem-based methods. Further experiments revealed no difference between visual and textual threat-based method in terms of actual effectiveness, while visual methods had better perception over textual ones.
- **A theoretical model** that extends MEM and hypotheses different characteristics of SRA methods which determine methods' actual efficacy and perceived efficacy. The model suggests that i) having a *clear process* supporting main steps of SRA and ii) availability of *visual summary* positively impact method's perceived ease of use, while iii) *modeling support* and iv) *help in identification of threats and security controls* by making available the catalogues of threats and security controls may increase method's perceived usefulness. The last feature (catalogues) also may affect method's actual effectiveness.
- **A theory** explaining how different features of catalogues of threats and security controls contribute into an effective risk assessment process for novices and experts in either domain or security knowledge.
- **An empirical investigation** on the comprehensibility of the different risk modeling approaches and a theoretical explanation of the findings.

1.3 Thesis Structure

The thesis is structured to present the main contributions of the Ph.D. study. The rest of the dissertation is organized as follows:

Chapter 2 presents a theory that hypotheses different characteristics of SRA methods which determine methods' actual efficacy and perceived efficacy. We also propose a robust evaluation framework to compare SRA methods, to evaluate the quality of results of methods application with help of external industrial experts and to identify aspects that may affect application of these methods. Finally, we present the results of three controlled experiments that validated and refined the proposed theory.

This chapter has been partially published or will appear in:

- [55] – K. Labunets, F. Massacci, F. Paci and L.M.S. Tran. "An Experimental Comparison of Two Risk-Based Security Methods". In *Proceedings of the 7th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2013.
- [96] – R. Scandariato, F. Paci, K. Labunets, K. Yskout, F. Massacci and W. Joosen. "Empirical Assessment of Security Requirements and Architecture: Lessons Learned". In *Engineering Secure Future Internet Services and Systems*. Springer, 2014.
- [58] – K. Labunets, F. Massacci, F. Paci and R. Ruprai. "An experiment on comparing textual vs. visual industrial methods for security risk assessment". In *Proceedings of the 4th IEEE International Workshop on Empirical Requirements Engineering (EmpiRE) at the 22nd IEEE International Requirements Engineering Conference (RE)*. 2014.
- [57] – K. Labunets, F. Paci, F. Massacci, M. Ragosta and B. Solhaug. "A First Empirical Evaluation Framework for Security Risk Assessment Methods in the ATM Domain". In *Proceedings of the 4th SESAR Innovation Days (SIDs)*. 2014.
- K. Labunets, F. Paci and F. Massacci. "An Empirical Comparison of Security Risk Assessment Methods". *Working paper*. To be submitted in journal.

Chapter 3 focuses on aspects of the theoretical model proposed in the previous section, particularly on the use of supporting artifacts like catalogues of threats and security controls. We propose a theory explaining how different features of catalogues contribute in an effective risk assessment process for novices and experts in either domain or security knowledge. In particular, we (1) examine the role of catalogues in the actual and perceived efficacy of an SRA; (2) compare the results of SRA with catalogues by non-security experts and without catalogues by security experts; (3) assess the role of catalogues' features in the actual and perceived efficacy of an SRA.

This chapter has been partially published or will appear in:

- [16] – M. de Gramatica, K. Labunets, F. Massacci, F. Paci and A. Tedeschi. "The Role of Catalogues of Threats and Security Controls in Security Risk Assessment: An Empirical Study with ATM Professionals". In *Proceedings of the 21st International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ)*. 2015.
- [56] – K. Labunets, F. Massacci and F. Paci. "Which Security Catalogue Is Better for Novices?". In *Proceedings of the 5th IEEE International Workshop on Empirical Requirements Engineering (EmpiRE) at the 23rd IEEE International Requirements Engineering Conference (RE)*. 2015.

- K. Labunets, M. de Gramatica, F. Massacci, F. Paci, M. Ragosta and A. Tedeschi. "On the Effectiveness of Sourcing Knowledge from Catalogues in Security Risk Assessment". *Working paper*. To be submitted in journal.

Chapter 4 focuses on another aspect of the theoretical model proposed in Chapter 2, namely the representation of security risks. We investigate the comprehensibility of two types of risk model representation: graphical versus tabular. We present the results of two empirical studies and explain our findings by proposing a simple extension of Vessey's cognitive fit theory.

This chapter will appear in:

- K. Labunets, F. Massacci, F. Paci, S. Marczak, F. M. de Oliveira. "Model Comprehension in the Context of Security Risk Assessment: An Empirical Comparison of Tabular vs. Graphical Representations". *Working paper*. To be submitted in journal.

Chapter 5 recaps main contributions described in this dissertation and reveals the future work.

Chapter 2

An Empirical Comparison of Security Risk Assessment Methods

This chapter aims to address our research question $RQ1$ by proposing a robust evaluation framework to compare SRA methods, to evaluate the quality of results of methods application with help of external industry experts and to identify aspects that can affect the application of these methods. To address research question $RQ2$, we report on a pilot and three controlled experiments on evaluation of SRA methods following the proposed framework. The pilot study and the first experiment were conducted with MSc students and professionals to compare different classes of methods: threat-based, goal-based and problem-based methods. The second and third experiments were conducted with MSc students to compare the best methods from the first experiment: visual and textual threat-based methods. We measured actual effectiveness as the quality of the security controls identified by the participants, while perceived ease of use and perceived usefulness were measured through post-task questionnaires. The main finding is that threat-based methods have better perceived efficacy than goal- and problem-based methods. Among threat-based methods, visual methods have higher perceived efficacy because they have a clear process and a graphical representation for assets, threats and security controls. However, it is unclear if they are actually more effective than textual methods.

To address research question $RQ3$, based on the experimental results we propose a theory that hypotheses different characteristics of SRA methods which determine methods' actual efficacy and perceived efficacy.

2.1 Introduction

Security controls (sometimes also denoted as countermeasures or security requirements) are usually identified using SRA methods. They identify the target systems' assets, the threats to those assets and the risk associated with those threats. Security controls are

then identified to mitigate the realization of the threats.

Many SRA methods, frameworks and standards exist – ISO 27005, NIST 800-30, COBIT, SABSA, etc. – but they all face the same issue: the process looks easy on paper – but it can be a complex and daunting task.

Despite the crucial role that SRA plays in the identification of security controls, only few papers [51,55,86,97,116,117,119] investigated which method is more effective for the identification of threats and security measures and why it is so. Evaluating SRA methods is challenging because it includes a number of confounding variables: the type of training received (e.g. all papers on the ISACA journal report methods application by a method's expert), previous expertise (e.g. student vs. professionals is a key distinction here), time allocated to a task, and presence of two analysis steps (threats identification and security controls identification depends on each other).

In this chapter we propose an empirical framework for evaluating and comparing different classes of security requirements elicitation methods with respect to three constructs defined in Method Evaluation Model (MEM) [77]: *actual effectiveness*, *perceived ease of use*, and *perceived usefulness*. To compare actual effectiveness of methods, we measure number of threats and number of security measures identified by each methods while controlling for the quality of identified threats and controls (domain experts from industry and method designers evaluate the results of participants). Participants' perceived easy of use and perceived usefulness are measured by post-task questionnaires. We also used qualitative analysis (post-it notes and focus groups) to understand drivers of a preference for a method over another one.

We have applied the framework to conduct experiments with both types of participants, professionals and MSc students. The overall evaluation reported in this chapter consist of a pilot study with 36 professionals and 13 MSc students preliminary reported in [65] and three controlled experiments: Experiments 2.1-2.3.

Experiment 2.1 was conducted with 15 MSc students in Computer science and 25 professionals in IT Audit for Information Systems. The participants were divided in 15 groups composed by MSc students and professionals following a between-participant design. Each group applied a method belonging to one of methods' classes to identify threats and security controls for one of two real application scenarios from Smart Grid and Healthcare domains.

In order to find better explanation of these findings we conducted two additional experiments. The participants of Experiment 2.2 and 2.3 were correspondingly 28 and 29 MSc students enrolled to Security Engineering course at the University of Trento. The aim of these experiments were to compare the best methods from Experiment 1: visual and textual threat-based methods. The participants of Experiment 2.2 were divided in 15 groups following within-participant design. Each group applied both methods to four tasks of the Smart Grid application scenario. The participants of Experiment 2.3 worked

individually and applied both methods to two tasks of the same Smart Grid scenario.

Experiment 2.1 found that threat-based methods are better with respect to perceived ease of use than goal-based and problem-based methods. The qualitative analysis suggests that the existence of a *clearly defined process* to identify threats and security controls is the major driver over and above using diagrams or tables, tools or mathematical foundations.

Experiment 2.2 shows that, processes being equally well defined, the visual method is better in identification of threats than the textual one. In contrast, Experiment 2.3 shows that the textual method is better in threats identification that the visual one when controlling for the results' quality. In both experiments (2.2 and 2.3) no statistically significant preference was found over security controls albeit tabular-based methods were slightly better in the first experiment when controlling for the quality of the results. An interesting open question was the potential role of catalogues in improving the effectiveness and preference of methods.

The reminder of the chapter is structured as follows. In the next section we discuss the related work. Then, Section 2.3 introduces our research context and research questions. Section 2.4 presents our framework to run comparative evaluations. After, Section 2.6 discusses the lessons that we learnt from the pilot study. The core of the chapter reports the execution and results of the three controlled experiments (in Section 2.5), results of the quantitative analysis of reports and post-task questionnaires, and of the qualitative analysis of questionnaires' open questions, post-it notes and individual or focus group interviews (Section 2.7. Then, we summarize our findings (Section 2.8) and discussed the refined theoretical model of methods' success criteria. Section 2.9 discuss threats to validity. In 2.10 we conclude the chapter.

2.2 Background on Identification of Security Measures

There are many standards, practices, and methods available for addressing information security risks, which differ in terms of focus and process. Several surveys review security methods based on theoretical frameworks that specify criterion for comparison [13, 22, 85, 126]. For example, [85] conducted a review of 11 security requirements engineering methods. They classified security requirements engineering methods with respect to the concepts that the methods are based on: goal-based, model-based, problem-based, and process-based methods. In our work we adopted a classification of security methods similar to the one proposed by [85].

All methods have a very similar process illustrated in Figure 2.1:

- **Assets Identification**. The goal of this step is to arrive at a correct understanding of the target of analysis and to pinpoint the most important valuables to be protected.
- **Threats and Security Risks Identification**. The goal is to identify possible threat scenarios targeting assets and classify risks that they represent based on the

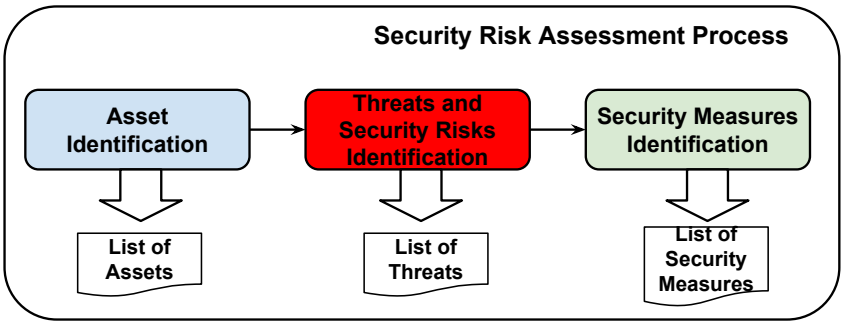


Figure 2.1: Overview of a Typical Security Risk Assessment Process

likelihood and impact of those threat scenarios.

- **Security Measures Identification**. The goal is to establish a list of measures to be achieved to address security risks.

However, specific instantiations are largely different in terms of focus, representation, and supporting material. Tables 2.1 and 2.2 compares examples of academic and industrial methods for SRA and compares them along these dimensions. In term of *focus* they can be divided as follows:

- *Threat-based methods* start their analysis from the identification of assets and related threats. Their steps are common to many SRA methods such as UK', Eurocontrol's or SESAR SecRAM. Academic methods in this realm are SREP, CORAS, LINDDUN.
- *Goal-based methods* focus on business goals and how they can be affected by threats and protected by corresponding security controls (or requirements, or measures) to them. This approach is typical of business methods such as COBIT or SABSA and academic methods like SECURE TROPOS, SI*.
- *Problem-based methods* use a problem-oriented notation proposed by Jackson [45] to describe the context for a target system and representing security controls as constraints on functional requirements. Examples of these methods can be found among academic methods: SECURITY ARGUMENTATION and ABUSE FRAMES.

With respect to *representation*, we can divide the selected methods in

- *Textual methods* that use text to describe assets, threats scenarios and security measures. SREP, SESAR and EUROCONTOL's SecRAM are examples of textual methods because they use tables to document the results of the execution of each step.
- *Visual methods* use a graphical representation to specify assets, threats scenarios and security measures. For example, CORAS uses diagrams to represent assets, threats and security measures while SECURE TROPOS uses goal models.
- *Hybrid methods* use both graphical and textual representations to express assets, threats and security measures. LINDDUN uses trees to represent threat scenarios but tables to map elements of the system under analysis to the threat scenarios. Similarly,

Table 2.1: Examples of Academic Methods

Method	Focus	Representation	Supporting Artifacts	Ref.
CORAS	Threat	Visual (Diagrams)	CORAS tool: Software to draw CORAS Diagrams	[61]
LINDDUN	Threat	Hybrid (Data Flow Models + Threat Tree Patterns + Tables)	Threat Tree Catalogue of Privacy Threats Scenarios	[18]
SEC. TRO.	Goal	Visual (Goal Models)	SecTro tool: Software to draw goal models	[82]
SEC. ARG.	Problem	Hybrid (Problem and Argument Diagrams + Argument Textual Formalization)	OpenPF tool: Software to create problem and argument diagrams, and to reason on arguments	[39]
SREP	Threat	Textual (Tables)	Security Resources Repository: Catalogue of Assets, Threats, and Security Requirements	[72]

Table 2.2: Examples of Industrial Standards and Methods

Method	Focus	Representation	Supporting Artifacts	Reference
BSI 100-2 IT-Grundschutz Methodology (DE)	Threat	Textual	BSI 100-4 IT-Grundschutz Catalogues (include assets, threats, and safeguards)	www.bsi.bund.de
COBIT (ISACA)	Goal	Textual (Tables)	Reporting Tool based on MS Excel	www.isaca.org
EBIOS (FR)	Threat	Textual	EBIOS tool	www.club-ebios.org
EUROCONTROL SecRAM	Threat	Textual (Tables)	EATM Catalogues of Threats and Security Controls	www.eurocontrol.int
MAGERIT (ES)	Threat	Textual	Catalogues of asset classes, valuation dimensions, valuation criteria, typical threats, and safeguards to be considered.	[10]
SABSA (Deloitte)	Goal	Textual	Enterprise Security Architecture Tool	www.sabsa.org
SESAR SecRAM	Threat	Textual (Tables)	MS Excel template supporting SecRAM process; SESAR SecRAM Tool; EATM Catalogues of Threats and Security Controls	Deliverable SESAR WP16.02.03: ATM Security Risk Assessment Methodology. www.sesarju.eu
HMG IA Standard Numbers 1 & 2 (UK)	Threat	Textual (Tables)	Baseline Control Set (catalogue of security controls)	www.cesg.gov.uk

SECURITY ARGUMENTATION uses both graphical and textual representations for expressing security arguments.

Methods can be distinguished also based on *supporting artifacts* that they offer to analysts for the execution of SRA. CORAS, SECURE TROPOS and SECURITY ARGUMENT provide only a software tool to draw graphical models, while the methods based on tables like SESAR SecRAM or COBIT are supported by Excel spreadsheet template. Some methods also give analysts security knowledge captured in catalogues of threats and security controls, e.g. industrial EUROCONTROL's and SESAR SecRAM, or academic SREP and LINDDUN.

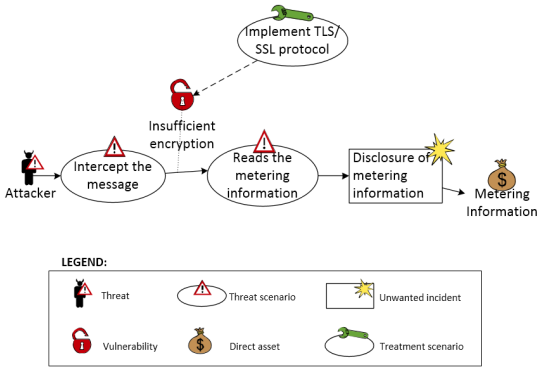
These methods are widely different with respect to focus, representation and supporting artifacts. They may not work equally well in identifying and mitigating security risks. Therefore, there is a need of conducting experimental comparisons to understand which methods are the best and why.

For further investigation we use the following academic methods: CORAS, SECURE TROPOS, SREP, LINDDUN and SECURITY ARGUMENTATION. Table 2.3 provides more details about them. The methods were selected based on (i) scientific profile (visibil-

Table 2.3: Methods' Overview

Method	Institution	Description
CORAS	SINTEF	Model-driven risk assessment method. CORAS is used by SINTEF for the industrial consultancies.
SREP	University of Castilla-La Mancha	Risk-driven and asset-based method to model and analyze security controls. This method is used by CMU Software Engineering Institute in their tutorials.
LINDDUN	Katholieke Universiteit Leuven	Risk-based method to elicit privacy requirements. This is based on STRIDE.
SEC. TRO.	University of East London	Goal-based method for modeling, and reasoning on security controls. The method supports capturing, analysis and reasoning of security controls from the early stages of the development process.
SEC. ARG.	Open University	Problem-based framework for security controls elicitation and analysis

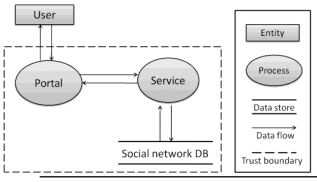
ity, citations, etc.), and (ii) availability of the scientists proposing them to hold a tutorial for our experiments. Criteria (ii) was important to avoid bias due to a lower quality training or a misinterpretation of method's key aspects. The restriction to academic methods in this phase was mostly due to financial reasons: training for SABSA by a SABSA specialist would cost almost 3000 euro per participant. Figure 2.2 shows example of artifacts produced during the application of the selected methods.



(a) CORAS - threat diagram

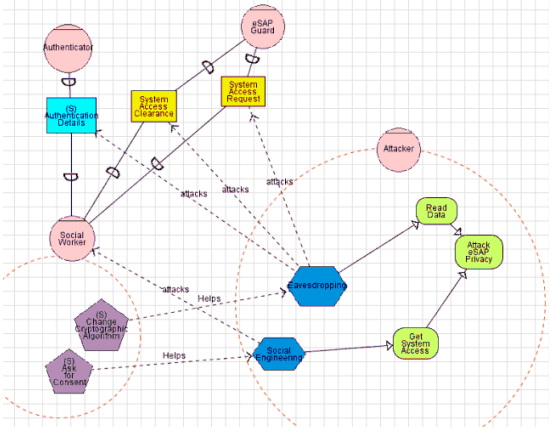
Name of Misuse Case: Spoof of information		
ID 1		
Summary: the attacker gains access to the message exchange between the SM and SNN and disclose the secret exchange of information		
Probability: Frequent		
Preconditions: 1) The attacker have access to the communication channel between SM and SNN		
User Interactions	Misuser interactions	System Interaction
The SM sends the information about power consumption		
	The attacker reads the information	
		The SSN receives the information without knowing that someone have read the message
Postconditions: 1) The attacker knows personal information about the power consumption of the customer		

(b) SREP - threat specification using misuse cases

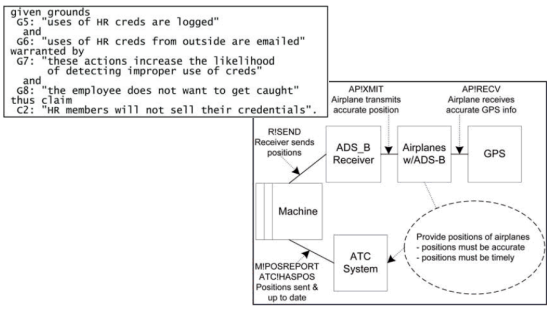


	Threat target	L	I	N	D	D	U	N
Data Store	Social network DB	×	×	×	×	×		×
Data Flow	User data stream (user – portal)	×	×	×	×	×		×
	Service data stream (portal – service)	×	×	×	×	×		×
	DB data stream (service – DB)	×	×	×	×	×		×
Process	Portal	×	×	×	×	×		×
	Social network service	×	×	×	×	×		×
Entity	User	×	×				×	

(c) LINDDUN - Data Flow Diagrams (DFD) and tables



(d) SECURE TROPOS - diagrams



(e) SECURITY ARGUMENTATION - textual description and diagram

Figure 2.2: Examples of methods' artifacts

2.3 Research Objectives and Theory

The goal of our experiments, according to the goal/question/metric template [4] is to analyze different SRA methods with the purpose of carrying out a comparative evaluation of these methods with respect to how successful they are in identifying threats and security controls from the point of view of security analysts.

In order to evaluate methods, it is first necessary to identify dimensions to measure success. Therefore, we have elaborated preliminary theory for evaluating SRA methods' success that is illustrated in Figure 2.3: the figure shows primary constructs and causal relationships between them. Some of the constructs reflect characteristics of SRA methods: *process/focus*, *representation*, *supporting artifacts*. The process/focus includes features like the fact that a method provides *clear guidelines* on how to conduct SRA process or the fact that the process starts from the identification of business goals (*goal-based* or assets and threats (*threats-based*). The representation of assets, threats and security measures can be either *visual*, *textual* or a mix of the two. The supporting artifacts are of different kind including *tool support* to document the results of SRA or *catalogues of threats and security controls*.

In addition, to allow comparison with the literature, we have considered constructs from MEM [77], which is a theoretical framework that incorporates constructs to evaluate methods' success. In particular, we considered *Actual Efficacy*, *Perceived Efficacy* and *Intention to Use (ITU)*. Actual Efficacy is the degree to which a method achieves its objectives (*Actual Effectiveness*) and is free of effort (*Actual Efficiency)*. Perceived efficacy is the degree to which person believes that a method achieves its intended objectives (*Perceived Usefulness (PU)*) and using it is free of effort (*Perceived Easy of Use (PEOU)*). *ITU* is the extent to which a person intends to use a particular method.

We also hypothesize that methods' characteristics like *process/focus*, *representation* and *supporting artifacts* determine method's actual efficiency and perceived efficacy. Therefore we can formulate our research questions in what follows.

RQ2.1 *Is actual efficacy significantly different between methods?*
Such research question is generic in nature and does not account for a specific nature of SRA. Several works [51, 59, 102] and a number of preliminary interviews with security experts has shown that there are two separate tasks which we must take into account: *a*) identification of threats or risks to assets and *b*) identification of security measures or controls to mitigate the risks. These tasks are located at different levels of the creativity spectrum: threats analysis requires thinking out of the box to anticipate attackers' behavior; identification of security measures requires a systematic review of threats to make sure they have been adequately mitigated. We have thus split research question *RQ*2.1 on actual efficacy into corresponding null-hypotheses for statistical hypothesis testing, one for threats analysis and one for controls identification.

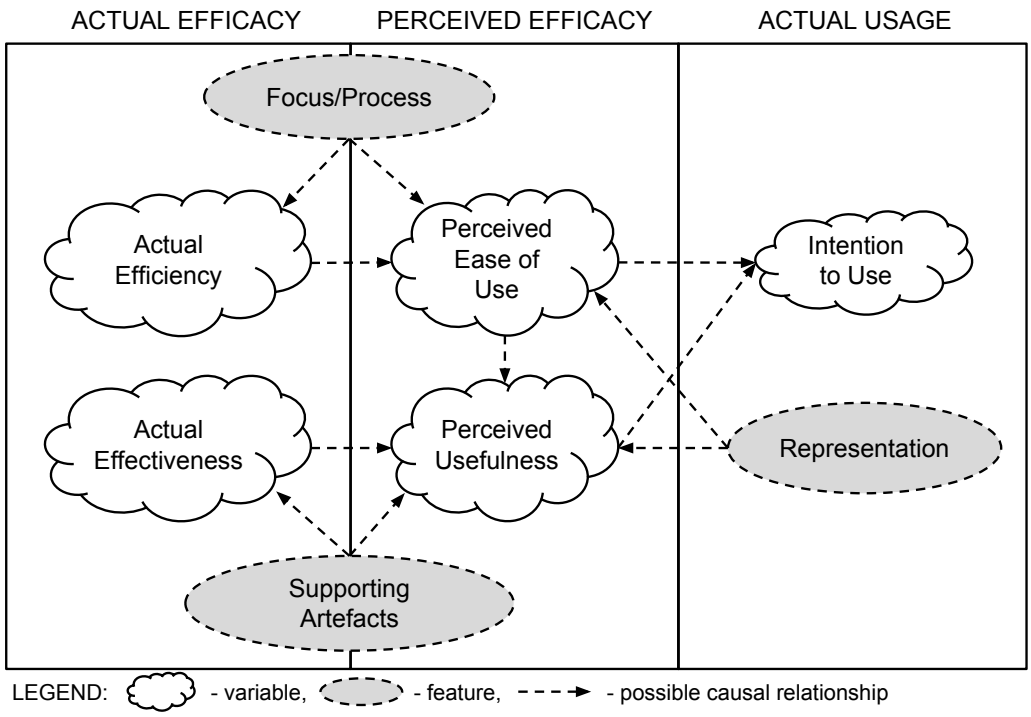


Figure 2.3: A Preliminary Model of Success Criteria for Security Risk Assessment Methods

RQ2.2 *Is participants' perceived efficacy significantly different between methods?*

RQ2.2.1 *Is participants' PEOU significantly different between methods?*

RQ2.2.2 *Is participants' PU significantly different between methods?*

RQ2.3 *Is participants' ITU significantly different between methods?*

The purpose of the research questions RQ2.2 and RQ2.3 is to evaluate methods' success with respect to the core concepts of MEM that were inherited from Technology Acceptance Model (TAM) [14]. Both approaches are widely used in the literature. [64] used these constructs to compare the perception of model-driven, model-based and code-centric software development methods. Similarly, [51] used TAM to compare perception of attack trees and misuse cases by industrial professionals. [112] and [83] applied MEM framework to evaluate methods that their proposed, respectively, in Security and Business Process Management fields. Moreover, TAM is widely adopted in Information System (IS) literature. Recently it was used to study IS acceptance in airline and banking domain [54] and adoption of mobile [29] and e-government IS [80].

RQ2.4 *What* qualitative drivers *may explain why a method is "better" than another one?*

One of the aspects that is not considered by MEM for methods' success is *how to determine key drivers* behind different PEOU and PU. This is what practitioners would actually want to know in order to select a method. It is also of interest for academic and industry

method's designers: how to improve one's own method. In these terms, the work by [93] is close to ours: the authors proposed cloud computing adoption model that refines TAM and Diffusion of Innovation theories and covers some domain and organization specific factors influencing technology adoption (e.g., complexity, compatibility, or infrastructure factors). The lack for similar study for SRA motivates the investigation behind our final research question $RQ2.4$.

2.4 Experimental Framework

This section describes the experimental framework to conduct controlled experiments for evaluating and comparing SRA methods. This should improve repeatability and comparability of results.

Conceptually the framework is divided in two parallel streams that are merged in time as shown in Figure 2.4: i) *Execution Stream* is the actual execution of an experiment in which methods are applied and its results are produced and validated; ii) *Measurement Stream* gathers quantitative and qualitative data that will be used to evaluate methods.

Each of the streams is divided into three phases:

- **Training**. The application scenario description is administered to participants by either an individual reading or introductory tutorial. Then, a frontal-training phase in which method's designer(s) introduce the method(s) to be used for SRA through a step by step tutorial.
- **Application**. The participants (individually or in groups) apply the assigned method to the scenario.
- **Evaluation**. The participants deliver the report documenting methods' outcomes. Several evaluators independently assess the quality of reports, providing marks and comments. The participants provide feedback on their experience with methods.

Four types of actors are necessary to implement this framework (besides researchers): *method designers*, *domain experts*, and *participants*. Method designers are the methods' inventors. Their main responsibility is to train participants in the method and to answer participants' questions during the Application phase. They evaluate group reports to determine if the method has been applied correctly. Domain experts are usually partners from industry who introduce the application scenario to participants. They evaluate the quality of security controls produced by each group of participants. They are also available during the Application phase to answers all possible questions that participants may raise during analysis. Participants have to identify threats and security controls for an application scenario using the assigned method. In the literature methods' designers are often the same people as the one that execute experiments. It introduces bias in methods' evaluation.

Tables 2.4a and 2.4b compare the steps of our experimental framework with the steps

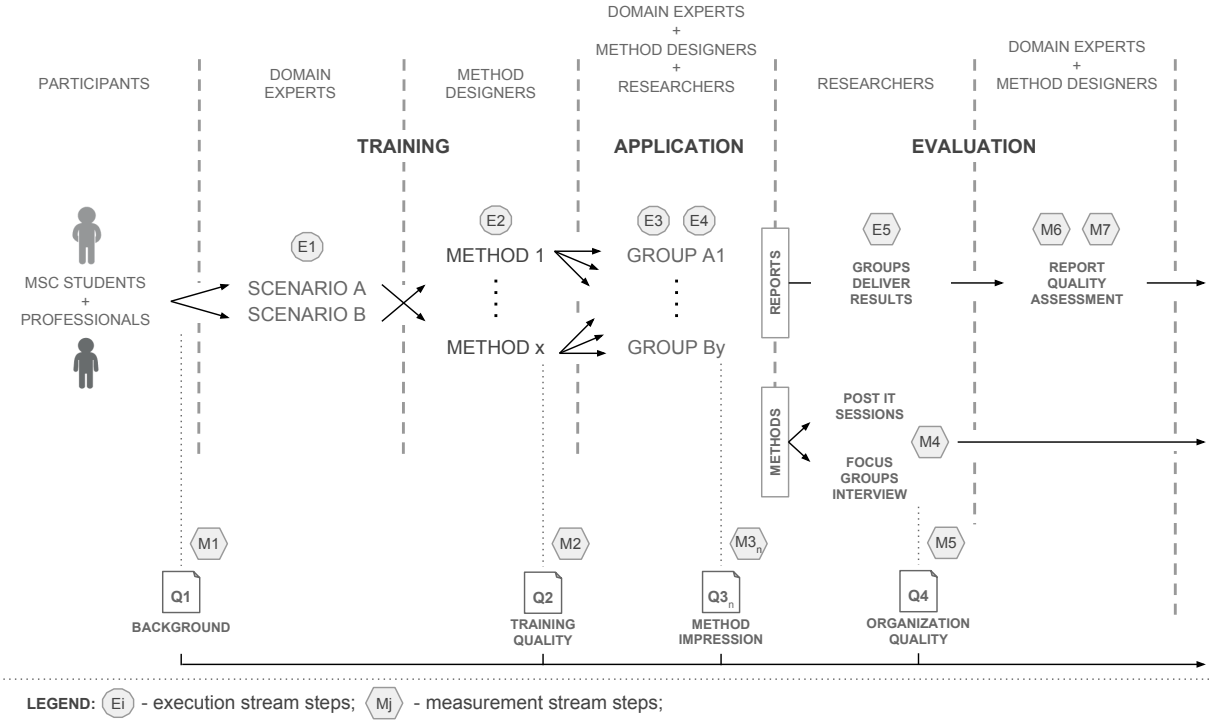


Figure 2.4: Experimental Framework

of the protocols used in the literature. We discussed these works in details below along with the description of framework steps.

2.4.1 Execution Stream

Training. Table 2.4a shows that half of the works reported that participants received training on the application scenario ($E1$). [97] provided participants with a significant training on the method and application domain scenario during three lectures of 2.5 hours, while [51] provided participants with a 90 minutes introduction about experimental goals and two application cases and made available to participants the materials for both application systems starting 4 days before the beginning of the experiment. At the same time, most of the works reported that participants were trained on methods. [86] provided only a brief 10 minutes training for each approach because they are easy to learn and no training on application cases as they were simple enough. [118] provided participants with an introduction to methods and description of an application case, but the authors did not specify the amount and format of the training (e.g. presentation by an expert or handout materials).

The first step ($E1$) targets the bias that might be introduced by previous knowledge of participants on a scenario. The "domain expert" provides to participants a uniform focus and target. It maybe his personal focus, but it is nonetheless the same for all participants.

Table 2.4: Framework's Steps in the Literature

(a) Execution Steps

Paper	$E1$: Scenario Tutorial	$E2$: Method Tutorial	$E3$: Application	$E4$: Feedback	$E5$: Final Report
Taubenberger et al. [123]			x		x
Opdahl et al. [86]		x	x		x
Massacci and Paci [65]	x	x	x	x	x
Labunets et al. [55]	x	x	x	x	x
Labunets et al. [58]	x	x	x	x	x
Scandariato et al. [97]	x	x	x		x
Stålhane and Sindre [115]		x	x		x
Stålhane and Sindre [116]		x	x		x
Stålhane et al. [119]		x	x		x
Stålhane and Sindre [118]	x	x	x		x
Karpati et al. [51]	x		x		x

(b) Measurements Steps

Paper	$M1$: Background	$M3$: Method Perception	$M4$: Focus groups	$M6$: Method Evaluation	$M7$: Results Evaluation
Taubenberger et al. [123]					
Opdahl et al. [86]		x			
Massacci and Paci [65]	x	x	x	x	
Labunets et al. [55]	x	x	x	x	x
Labunets et al. [58]	x	x	x	x	x
Scandariato et al. [97]	x	x			x
Stålhane and Sindre [115]	x	x			
Stålhane and Sindre [116]	x	x			
Stålhane et al. [119]	x	x			
Stålhane and Sindre [118]	x	x			
Karpati et al. [51]	x	x			

The rationale of the second step ($E2$) is to limit the implicit bias that might be introduced by having to train participants into one's own method and a competitor's one.

Application. We tried to make the application session ($E3_n$) last at least 16 hours of work. We believe this is necessary to fully exercise the method. [97] reported that their participants spent around 4 full days to model threats with STRIDE methodology. In contrast, several papers in the IS and Requirements Engineering (RE) literature limited method's application to less than 2 hours. For example, [123] reported a controlled experiment where 55 security professionals have 30 minutes to conduct SRA. The participants of the experiment reported by [86] had only 30 minutes to find threats using one of two techniques. In the replication [51] professionals spent around 2 hours to complete the task. The other works [118, 119] also reported the use of step $E3_n$.

Group presentations ($E4_n$) are essential to capture a phenomenon present in reality and namely domain expert feedback and internal presentation. They might indeed bias analysis, as participants will adjust their work along the received feedback. Yet, this is precisely what happens in reality. We considered the benefit for external validity greater than the threat to conclusion validity. Only few works reported the use of both $E3_n$ and $E4_n$ steps [97, 118].

Evaluation of the results. This step *(E5)* is widely adopted in the literature, as it is essential for capturing the results of any process application. All works listed in Table 2.4a mentioned this step as a part of their experimental procedure. For example, in [86] participants were asked to deliver threats in misuse-case or attack tree diagram format with a brief textual explanation of each threat if necessary.

2.4.2 Measurement Stream

Experiment's Outcome. Before method's application, participants are administered a questionnaire to collect information about their level of expertise in requirement engineering, security and on other methods they may know *($M1$)* and a post-training questionnaire to determine their initial perception of the methods and the quality of the tutorials *($M2$)*. Step $M1$ is common for empirical studies and used to control possible effects of participants' background on experimental results. However, [123] and [86] did not mention if they performed similar step. Step $M2$ is not widely used in similar experiments. Afterwards, at step $M5$ participants are requested to answer a post-task questionnaire about the quality of experiment organization (Q4). Similar to $M2$, this information is used to control a possible effect of the experiment settings on results.

Participant's Quantitative Evaluation. After each application session participants are requested *($M3_n$)* to answer a post-task questionnaire ($Q3_n$) about their perception of the methods. This step is a key one for empirical studies aiming to compare different methods. In particular, we adopted this post-task questionnaire from MEM. Almost all works listed in Table 2.4b also mentioned that they collected participants' perception of the methods after the application.

Participant's Qualitative Evaluation. The goal of the next step ($M4$) is to *collect* participants' perception and feedback on the methods through post-it notes sessions and focus group interviews. However, most of works do not conduct qualitative study (e.g. by mean of interviews with the participants) why a method is better than the others, but just collect participants' perception through questionnaires. After reports delivery, participants are divided in groups based on the assigned methods. They are involved in focus groups interviews where they are asked questions on their perception of methods. A separate post-it note session is run with each group. In each session, participants are requested 1) to annotate on post-it notes 5 positive and 5 negative aspects of the applied method and 2) hang the post-it notes on a wall and group post-it notes that reports similar opinions about method's aspects. Once grouped, the post-it notes have to be listed in order of importance.

Evaluation of the results by experts. The goal of this phase is to validate whether participants applied correctly the method and identified threats and security controls are specific for scenarios. First, ($M6$) groups' reports are evaluated by method designers.

Table 2.5: Types of Data Collected by the Framework

Data source	Description	Use	Type
Questionnaires	Include questions on participants' knowledge of IT security, risk assessment and their evaluation of the methods' aspects	Demographics, PEOU, PU, ITU	Quantitative
Audio/Video Recordings	Capture application of methods by participants and record focus groups interviews	PEOU, PU	Qualitative
Post-it Notes	List positive and negative aspects about the methods	Drivers, PEOU, PU	Qualitative
Focus Group Transcripts	Report the discussion on the methods' application between participants and observers	Drivers, PEOU, PU, ITU	Qualitative
Group Presentations	Participants summarize the results of method's application	Actual Effectiveness	Qualitative
Final Reports	Describe in detail how participants have identified the security controls following the method	Actual Effectiveness	Quantitative

They evaluate the quality of method's application. The level of quality is on a four item scale: *Unclear* (1), *Generic* (2), *Partial* (3) and *Total* (4). After, ($M7$) the group reports are evaluated by *domain experts*. Domain experts assess the quality of identified threats and security controls. The level of quality is on a four item scale: *Unclear* (1), *Generic* (2), *Specific* (3) and *Valuable* (4). Table 2.4b shows that only one work [97] evaluated the quality of the results produced by participants like we do in our framework ($M6$).

2.4.3 Rationale

Steps $M6$ and $M7$ address two issues that may affects both conclusion and construct validity. Literature opinion varies on whether the quality of results should be evaluated by some independent expert. Some authors [23] argue that it is not necessary, other papers don't mention whether this steps have been performed [51,118], other people would deem it essential and practitioners put it "If the quality of your risk assessment doesn't matter then any method works well." Any method can be effective if it does not need to deliver useful results for a third party (hence the evaluation by a domain expert). It can also be properly easy to use if participants do not follow it (hence the evaluation by a method designer). It is important to show which method is better in delivering not just results but good ones [59, Ch. 3]: *"the security risk assessment report is expected to contain adequate and relevant evidence to support its findings, clear and relevant recommendations, and clear compliance results for relevant information security regulations."*

In order to assess actual effectiveness ($RQ2.1$), final reports delivered by groups were evaluated by domain and methods' experts $M6$ and $M7$. In order to lessen the load of domain experts, researchers count the number of threats and security controls in the reports and provide the list of threats and controls to the experts.

To evaluate perceived easy of use ($RQ2.2.1$), perceived usefulness ($RQ2.2.2$) and intention to use ($RQ2.3$) we look at the answers on questionnaires distributed at step $M3_n$.

Table 2.5 summarizes types of data that we collect and how they measure different aspects (actual effectiveness, PEOU, PU and ITU).

Table 2.6: Statistical Tests Selection

Comparison Type	**Interval/Ratio (Normality is assumed)**	**Interval/Ratio (Normality is not assumed), Ordinal**
2 paired groups	Paired t-test	Wilcoxon test
2 unpaired groups	Unpaired t-test	Mann-Whitney (MW) test
3+ matched groups	Repeated-measures ANOVA	Friedman test
3+ unmatched groups	ANOVA	Kruskal-Wallis (KW) test

To analyze final reports and questionnaires, we select statistical tests depending on the design type and the assumption on normal distribution of the samples. Table 2.6 (shortened version of the Table 37.1 from [40, Chap. 37]) gives an overview of how we select statistical tests in our empirical studies. Actual statistical tests being used will be reported in the sections describing each individual study. For all statistical tests we adopt 5% as a threshold of α (i.e. probability of committing Type-I error) [128].

Instead, to investigate responses to open questions about positive and negative aspects of the methods, we analyze post-it notes and transcripts of focus group interviews using coding [120] – a qualitative analysis method used in grounded theory. The analysis comprises the following steps:
Open coding. Each researcher analyzed open questions and post-it notes to identify codes that represent patterns related to methods' positive (negative) aspects.
Focused coding. The researchers worked in groups to code and classify the codes into iteratively emergent categories relevant for PEOU and PU.
Theoretical coding. The researchers identified core categories and relationships between them.

2.5 Summary of Experiments

Figure 2.5 provides a bird's eye view of the empirical studies and key components of the experiments. The pilot study debugged our experimental framework. Experiment 2.1 further polished the experimental framework and gave insights about what impacts effectiveness and perception of different classes of IT SRA methods. For experiments 2.2 and 2.3 we chose the methods of the class that were perceived as better methods by participants in Experiment 2.1 in order to further investigate the reasons behind their success. Table 2.7 compares the experiments. All experimental materials used by participants (i.e. scenarios description, methods tutorials) are available on-line[1].

Demographics Tables 2.8-2.10 report descriptive statistics about participants of the three experiments. We have spent a significant effort by incorporating professionals because having only students is known to be a major threat to external validity [89].

[1] `https://securitylab.disi.unitn.it/doku.php?id=validation_of_risk_and_security_requirements_methodologies`

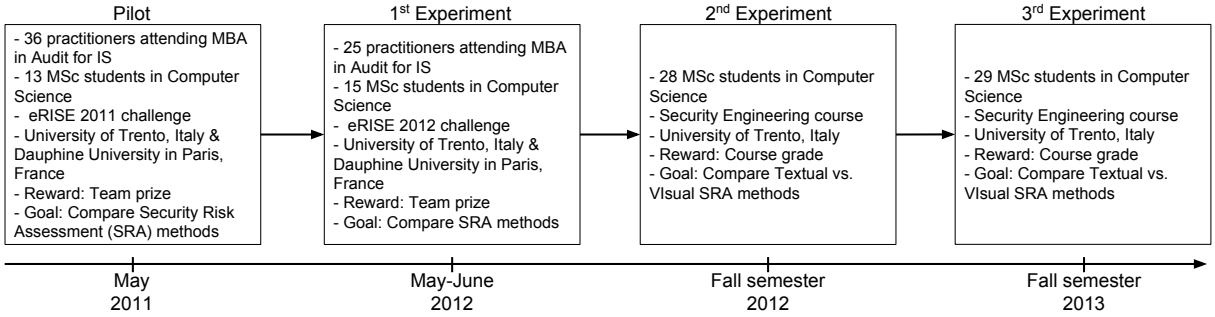


Figure 2.5: Empirical Studies Overview

Table 2.7: Summary of the Experiments

Setting	**Pilot (2011)**	**Exp. 2.1 (2012)**	**Exp. 2.2 (2013)**	**Exp. 2.3 (2014)**
Participants	36 professionals + 13 MSc students	25 professionals + 15 MSc students	28 MSc students	29 MSc students
Design Type	Between-participant	Between-participant	Within-participant	Within-participant
Methods	CORAS, LINDDUN, SEC. TRO., SI*, and SEC. ARG.	CORAS, LINDDUN, SEC. TRO., SREP, and SEC. ARG.	CORAS and SREP	CORAS and EUROCONTROL SecRAM
Scenarios	Healthcare Collaborative Network	Smart Grid and e-Health	Smart Grid, different tasks	Smart Grid, different tasks
Variables	N/A	PEOU, PU	Actual Effect., PEOU, PU, ITU	Actual Effect., PEOU, PU, ITU

Experiment 2.1: The experiment involved 15 MSc students in Computer Science from the University of Trento and 25 professionals who were attending a part-time MBA in Audit for Enterprise Information System at Paris Dauphine University where students spend half of the week working in consulting companies from different domains like Management Consulting Services and Audit (PwC, Accenture plc), Oil and Gas industry (Total S.A.), Pharmaceutics (Sanofi S.A.), Telecommunications (SFR), and Banking (Banque de France, Exane, RCI Banque). Participants were divided in 15 groups composed by one MSc student and one or two professionals. A significant fraction (30%) of participants reported that they worked specifically on security/privacy projects. The rest of participants (40%) reported no information about their work experience.

Experiment 2.2: Participants for the experiment were recruited among MSc students enrolled in a Security Engineering course at the University of Trento. The experiment involved 28 MSc students. Some participants (18%) reported that they were involved in security/privacy activities. Majority of the participants (60%) reported that they had working experience while the remaining did not provide any information.

Experiment 2.3: The participants were 29 MSc students enrolled in a Security Engineering course at the University of Trento. Similar to Experiment 2.2, 18% of participants reported that they were involved in security/privacy activities. Most of participants (69%) reported that they had at least 2 years of working experience while the remaining reported no working experience.

Table 2.8: Overall Participants' Demographic Statistics – Experiment 2.1

Variable	Scale	Mean	Med.	Distribution
Age	Years	25.46	-	58.97% were 20-24 years, 33.33% were 25-29 years and 7.69% were 30-60 years old
Work experience	Years	2.97	-	17.5% not specified, 15% have less 1 year, 42.5% have 1-2 years, 25% have $>$ 2 years
Expertise in Security	1(Novice)-5(Expert)	2.43	2	15% novices, 42.5% beginners, 27.5% competent users, 15% proficient users, 0% experts
Expertise in RE/Modelling	1(Novice)-5(Expert)	2	2	45% novices, 25% beginners, 15% competent users, 15% proficient users

Table 2.9: Overall Participants' Demographic Statistics – Experiment 2.2

Variable	Scale	Means	Med.	Distribution
Age	Years	24.33	-	51.85% were 21-24 years, 44.44% were 25-29 years and 3.7% were 30-35 years old
Work experience	Years	2.76	-	39.3% not specified, 28.6% have $\leq$ 2 years, 32.14% have $>$ 2 years
Expertise in Security	1(Novice)-5(Expert)	2.18	2	25% novices, 39% beginners, 29% competent users, 7% proficient users
Expertise in RE/Modelling	1(Novice)-5(Expert)	2.57	3	29% novices, 18% beginners, 25% competent users, 25% proficient users, 4% experts

Table 2.10: Overall Participants' Demographic Statistics – Experiment 2.3

Variable	Scale	Means	Med.	Distribution
Age	Years	25.72	-	48% were 21-24 years; 41% were 25-29; 10% were 30-40
Work experience	Years	2.46	-	31% had no experience; 31% had $<$ 2 years; 28% had 3-5 years; 10% had $>$6 years
Expertise in Security	1(Novice)-5(Expert)	2.38	2	28% novices; 31% beginners; 21% competent users; 17% proficient users; 3% experts
Expertise in RE/Modelling	1(Novice)-5(Expert)	2.31	2	24% novices; 34% beginners; 28% competent users; 14% proficient users

Application Domain Selection To conduct our experiments we selected two different industrial application scenarios from Siemens and Atos Research:
E-Health. The application scenario by Siemens was related to the management of electronic healthcare records. The scenario focused on registering new patients in a clinic and includes assigning clinicians (doctors, nurses, etc.) to patients, reading and updating a record, retrieving patient information from external sources, and providing results of examinations and treatments to authorized externals entities.
Smart Grid. Atos Research proposed a scenario about Smart Grid which is an electricity network using information and communication technology to optimize the transmission and distribution of electricity from suppliers to consumers. In particular, the scenario focused on a smart meter which records consumption of electric energy and communicates daily this information back to the utility for monitoring and billing purposes.

In Experiment 2.2 the Smart Grid scenario was refined into a number of tasks. The tasks were Security Management (Mgmnt), Web Application/Database Security (WebApp/DB), Network/Telecommunication Security (Net/Teleco), and Mobile Security (Mo-

Table 2.11: Collected Data and Research Questions

Variable	**Exp. 1**	**Exp. 2**	**Exp. 3**
RQ2.1: Actual Effectiveness	Final reports: sec. controls ($E5$, $M6$ and $M7$)*	Final reports: threats & sec. controls ($E5$, $M6$, $M7$)	Final reports: threats & sec. controls ($E5$, $M6$, $M7$)
RQ2.2.1: PEOU	Post-task questionnaire ($M2$ and $M3_n$)	Post-task questionnaire ($M3_n$)	Post-task questionnaire ($M3_n$)
RQ2.2.2: PU	Post-task questionnaire ($M2$ and $M3_n$)	Post-task questionnaire ($M3_n$)	Post-task questionnaire ($M3_n$)
RQ2.3: ITU	N/A	Post-task questionnaire ($M3_n$)	Post-task questionnaire ($M3_n$)
RQ2.4: Qual. drivers	Post-it notes, focus group interviews ($M4$)	Individual interviews ($M4$)	Individual interviews ($M4$)

*Note: In parentheses we report steps of the framework that are involved in data collection

bile). For example, in the WebApp/DB task, groups had to identify application and database security threats like cross-site scripting or aggregation attacks and propose mitigations. In Experiment 2.3 we used only Network and WebApp/DB security tasks as participants were asked to work individually.

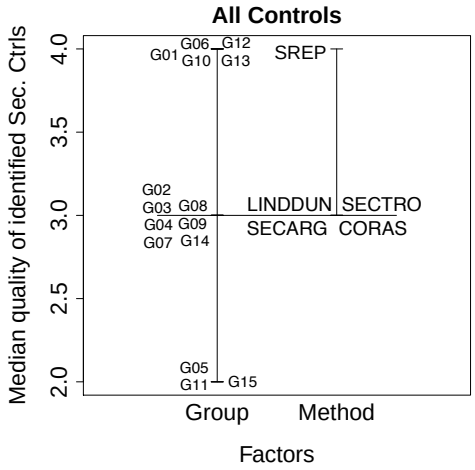
Collected Data In Experiment 2.1 we collected 15 methods applications. In Experiment 2.2 we collected 64 methods applications (16 groups $\times$ 4 tasks) and each group summarized the identified threats and security controls in a single report per group. In Experiment 2.3 participants were asked to work individually. In total, we collected 58 methods applications (29 participants $\times$ 2 tasks). In each of three experiments participants were asked to reply individually to a post-task questionnaire. Table 2.11 summarizes the data that were collected during the experiments.

2.6 Lessons Learned from the Pilot Study

The pilot study helped us to refine research questions and our experimental framework. The most important lessons are stated below. We stress that i) they mostly concern practitioners and ii) are often not reported from the literature. We speculate that our problems raised because experiments in the literature are usually run with students who tend to be compliant to instructor's guidelines.

Limit data collection. Too frequent measurements disrupt the natural flow of activities of an experienced group or could be perceived by participants as intrusive. The former introduces bias in the process and the latter leads to participants dropping from the experiments.

Avoid usage of previous knowledge. Practitioners have experience of "appropriate" security controls from past job assignments. If they think they have already seen the scenario, then they will deliver their pre-cooked security controls (without running through any method at all). This is a major threat to validity and it is very difficult to spot if one only collects the final requirements. So, our final report always includes a brief section



Only groups using SREP demonstrated better results. Unfortunately, the experiment does not allow to draw statistically significant conclusions on actual effectiveness due to the small sample (an average of 3 groups per method).

Figure 2.6: Median Quality of Identified Security Controls – Experiment 2.1

where participants are supposed to illustrate how the method allowed deriving the final controls.

Clarify importance of assigned method application. Practitioners (and students to some extent) focus on results. Some groups thought security controls mattered so they ditched the assigned method ("it was not working") and used a completely different method that they already knew. This can be detected by reading the section on the method, but the work of the group cannot be used; it is a data loss.

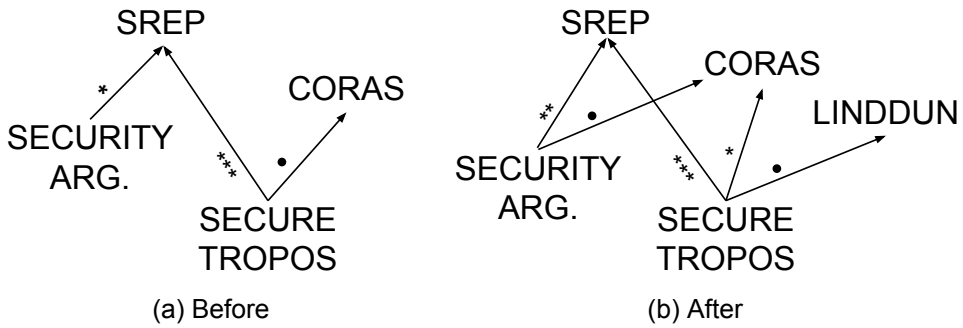
Have method designers and domain experts available. The presence of method's designers and domain experts during the Application phase allows participants to ask for additional information that may have not been provided during Training.

Beware of language issues. Most studies in the literature are mono-lingual and this aspect is overlooked, whereas the participants from our studies were of mixed nationalities. Care should be taken during focus groups sessions to misinterpret or lose feedback because participants do not feel confident to speak in English.

2.7 Results

We compared different SRA methods with respect to their actual efficacy that was measured in terms of number of identified threats and security controls and their quality. The results of Experiment 2.1 (see Figure 2.6) revealed that the textual methods helped groups to identify security controls of a better quality (median quality is 4) than other methods (median quality is 3).

The results of the post-task questionnaire that measured methods' perceived efficacy



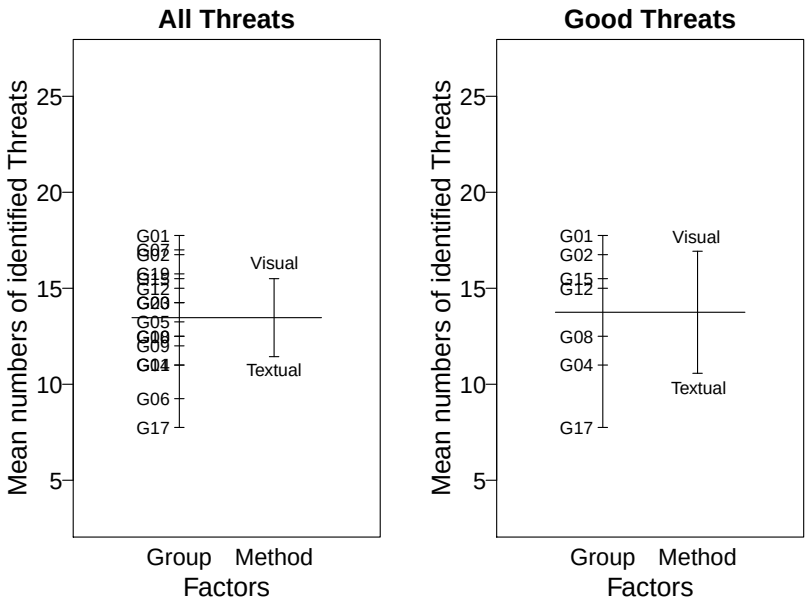
To compare methods we collected answers to all questions about perceived efficacy (both PEOU and PU) and ran a post-hoc MW test on the comparison (appropriately corrected for multiple tests) of methods X and Y: *Perceived efficacy of X is better than perceived efficacy of Y*. We draw an arrow from method X to method Y if method Y has a statistically significant higher perceived efficacy than method X. The vertical position has been spaced to reflect the relative level of the answers.
After the training, only the textual and visual threat-based methods are perceived as better than other non-threat-based methods. After a calendar month of remote application and almost two days full time of controlled application, all threat-based methods are perceived as better than others. Observations are statistically significant (post-hoc MW test).

Figure 2.7: Overall Assessment of Methods' Perception – Experiment 2.1

(see Figure 2.7) show that threat-based methods (SREP, CORAS and LINDDUN) were perceived by participants to be superior to other methods. Therefore, in consequent experiments we focused on two types of threat-based security methods, namely visual and textual.

2.7.1 Actual Efficacy

Experiment 2.2 showed that the visual method is more effective in identifying threats (on average 50% more threats) than the textual one for both good and all groups (see Figure 2.8), while the textual method was found to be more effective in identifying security controls (on average 20% more controls) for good controls and this is supported by the Friedman test ($p = 7.4 \cdot 10^{-3}$). The division on bad and good threats and security controls were done based on the assessment results by domain experts. Experiment 2.3 aimed to generalize the previous results and investigated different textual method. The results of the third experiment, in contrast to the second one, showed that the textual method is more effective in identifying threats (on average 40% more threats) for good threats (see Figure 2.9) but this result is not statistically significant. There is also no difference between two methods in identifying security controls.



The visual method performed better in threats identification in both cases: if we consider threats of any quality (on average 50% more threats) and if we apply quality filter and take only good ones. Observations are statistically significant in the number of threats of any quality (ANOVA test returned $p = 1.95 \cdot 10^{-4}$ and Friedman test returned $p = 8.9 \cdot 10^{-3}$) and good threats (Friedman test returned $p < 1 \cdot 10^{-3}$).

Figure 2.8: Numbers of Identified Threats by Quality – Experiment 2.2

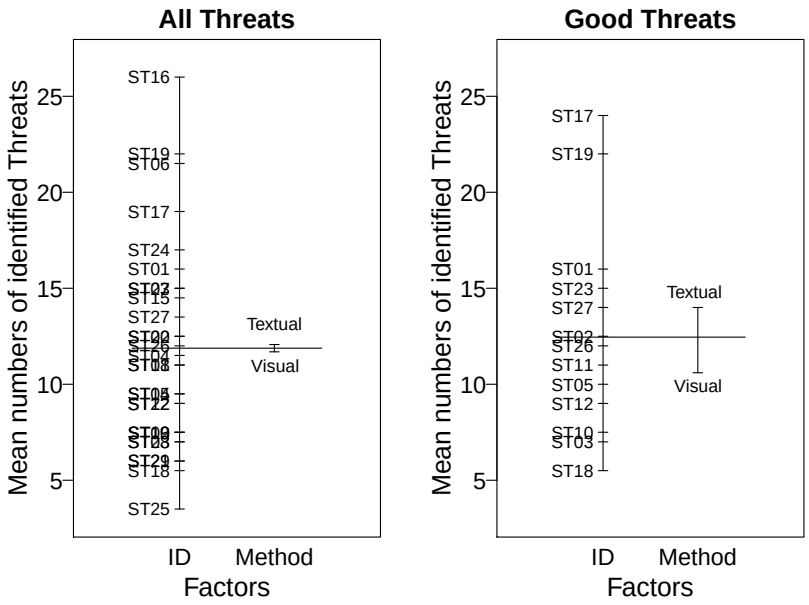
2.7.2 Perception

The results of questionnaire analysis (see Table 2.12) show that the visual method is better than the textual one with respect to each perception variable (PEOU, PU, ITU) across all participants but this is not statistically significant. Good participants demonstrate a statistically significant preference for the visual method for ITU, and a small but not statistically significant preference for the visual method with respect to PEOU and PU. Table 2.13 presents the results of Experiment 2.3 and shows that the visual method is better than the textual one with respect to each perception variable (PEOU, PU, ITU) across all and good participants and this is statistically significant.

2.7.3 Qualitative analysis

Experiment 2.1: From responses to open questions, post-it notes and transcript of focus group interviews we coded 159 positive and 139 negative statements on PEOU, and 38 positive and 18 negative statements on PU. The results are detailed in Section 6.1 in Tables 6.5a and 6.5b.

CORAS, SREP and LINDUN had most of positive comments related to PEOU which were respectively 40, 41, and 31. All other methods had less than 30 positive comments.



The textual method is found to be more effective in identifying threats than the visual one. But the results of the statistical tests did not confirm this for both all threats (Friedman test returned $p\text{-}value = 0.57$) and good threats (Skillings–Mack test returned $p\text{-}value = 0.17$).

Figure 2.9: Numbers of Identified Threats by Quality – Experiment 2.3

Table 2.12: Participants' Perception by Variables and Quality of Results – Experiment 2.2

	All participants				**Good participants**		
Variable	**Median**		Z_W	Z_{MW}	**Median**		Z_{MW}
	Textual	**Visual**			**Textual**	**Visual**	
PEOU	3	3	-0.57	-0.9	3.5	4	-1.82 •
PU	3	4	-1.35	-0.96	3	4	-1.69 •
ITU	3	3	-0.86	-1.01	3	4	-2.59 *

Table 2.13: Participants' Perception by Variables and Quality of Results – Experiment 2.3

	All participants				**Good participants**		
Variable	**Median**		Z_W	Z_{MW}	**Median**		Z_{MW}
	Textual	Visual			Textual	Visual	
PEOU	3	4	-6.51 ***	-6.16 ***	2.5	4	-4.19 ***
PU	3	4	-4.82 ***	-4.56 ***	3	4	-3.88 ***
ITU	3	4	-3.57 ***	-3.67 ***	3	4	-2.94 ***

Tables report participants' responses to questions aggregated by perception variable (PEOU, PU, ITU), the median of responses by all and by good participants (the one who were part of groups that produced good quality threats and security controls based on experts' assessment), and the level of statistical significance based on the p-value returned by the Wilcoxon test for the paired comparison (all participants) and the MW test for both all and good participants. Note: • - p-value <0.1, * - p-value <0.05, ** - p-value <0.01, *** - p-value <0.001

Negative comments on PEOU where distributed among various methods. Each of them faring around 30 statements except LINNDUN which only had 16 negative statements. There were very few comments on PU either positive or negative (less than 10 per method).

In summary, the following elements *a*) *clear process*, *b*) *easy to understand*, and *c*) *visual summary* are the main aspects impacting PEOU of studied methods, and *a*) *modeling support*, and *b*) *security/privacy specificity* are the key aspects influencing methods' PU. For example, the presence (resp. absence) of a clear process was one of the most frequent causes offered by participants to describe their reason for methods' perceived ease of use. It accounted for 31% of positive statements (resp. 21% of negative statements) over the total number of recorded statements. For CORAS (40% positive statements) and LINDDUN (29% of positive statements and no negative one) having a clear process positively affects their PEOU. Here are some examples: "For me it was very clear steps from the first till the last one." (CORAS); "The process is very clear and it is easy to understand the method." (LINDDUN); "The process is not so technical, so it is easy to understand." (SREP). For other methods participants stressed that methods were convoluted or just not clear: "I think the process of the method is heavy, slow, complex to follow." (SECURITY ARGUMENTATION).

This provides a clear explanation for the measured perceived superiority of threat-based methods (SREP, CORAS and LINDDUN) over other methods. In fact, the former methods have clear process to follow.

Some participants pointed that having a visual summary was also important. Both CORAS and SECURE TROPOS have a visual representation language and participants appreciated that: 15 people mentioned it for CORAS and 4 mentioned it for SECURE TROPOS. SECURE TROPOS has also a richer modeling language and 7 participants explicitly mentioned it ("I liked the fact that it helps you to model the use case that you are treating."). Yet, this was *not* enough to revert the judgment on the ambiguity of process, and hence the less positive appraisal of SECURE TROPOS over CORAS.

In summary, our analysis shows that the main driver is process clarity, while other aspects are second order drivers.

Experiment 2.2: We analyzed transcripts of individual interviews and coded 80 positive and 53 negative statements on PEOU, and 85 positive and 20 negative statements on PU. Tables 6.12a and 6.12b in Section 6.2 detail results.

Visual method had most (both positive and negative) of statements related to PEOU: 53 out of 80 positive and 33 out of 53 negative statements out. With respect to PU textual method had most of positive statements (46 out of 85 statements) while visual method had more negative statements (19 out of 20 statements).

The results of qualitative analysis show that *a*) *clear process*, and *b*) *visual summary* are the main aspects impacting methods' PEOU, while *a*) *complexity of visual summary*, and *b*) *help in identifying threats and security controls* are the main aspects influencing methods' PU. Like in Experiment 2.1 participants of Experiment 2.2 reported a clear process as one of the main aspects that describes methods' perceived ease of use. They made 35% of positive statements (resp. 25% of negative statements) over the total number

of PEOU statements. For CORAS (23% of positive statements) having a clear process has a positive effect on its PEOU. For example, "steps are very clear." But there is no consensus about clear process of SREP because participants made 59% of positive and 55% negative statements about this aspect. Here are some examples: "Well defined steps. Clear process to follow" and "steps are not well explained."

Similar to Experiment 2.1, another important PEOU aspect reported by participants is availability of visual summary. About 45% of positive statements in CORAS were made by participants in relation to this aspect. A typical statement was: "Diagrams are useful. You have an overview of the possible threat scenarios and you can find links among the scenarios".

Experiment 2.3: In Experiment 2.3 we analyzed transcripts of individual interviews and coded 161 positive and 212 negative statements on PEOU, and 107 positive and 63 negative statements on PU. Table 6.16 in Section 6.3 detail results.

Visual method had most (both positive and negative) of statements related to PEOU: 121 out of 161 positive and 115 out of 212 negative statements out. With respect to PU textual method had most of negative statements (37 out of 63 statements) while visual method had more positive statements (71 out of 107 statements).

Experiment 2.3 supports the findings of the qualitative analysis of Experiment 2.2 both for PEOU and PU aspects. Similar to Experiments 2.1 and 2.2, participants of Experiment 2.3 supported that a clear process is among the main aspects defining methods' PEOU: 12% of positive statements (resp. 9% of negative statements) over the total number of PEOU statements. For CORAS 23% of positive statements were made about this aspect: "the advantages of CORAS is very clear structure". For the textual method still there is no agreement on clear process of SecRAM: participants made 45% of positive and 29% of negative statements about it. For example, "the steps are very clear" and "the steps and even the methodology was not really quite clear".

Another important PEOU aspect is having a visual summary: participants made 36% of positive statements about this aspect for CORAS. Here some examples: "there are many summary diagrams which are useful to summarize what has been done" and "the advantage is the visualization".

The results of Experiments 2.2 and 2.3 can explain participants' perceived preference of the visual method over the textual one. In conclusion, the analysis results support the findings of Experiment 2.1 and show the importance of clear process and visual summary as key drivers.

2.7.4 Correlation Analysis

Based on the results of Experiment 2.2 we conducted a correlation analysis between actual effectiveness, PEOU, PU and ITU to validate relations proposed by MEM. Our data have

Table 2.14: Summary of Empirical Results

Note: in *italics* reported statistically significant results, in normal text reported results with 10% significance level unless explicitly mentioned that there is no statistically significance.

RQ & Concept	**Results**		
	Experiment 2.1	**Experiment 2.2**	**Experiment 2.3**
*RQ*2.1 Actual effectiveness	OPEN (Not enough data)	a) *More threats were found with visual method than with textual method.* b) Slightly more security controls were found with textual method than with the visual one BUT the difference is not statistically significant.	a) Slightly more threats were found with textual method than with visual one BUT the difference is not statistically significant. b) No difference in the number of identified security controls.
*RQ*2.2.1 PEOU	*PEOU is higher for threat-based methods with statistical significance.*	Threat-based visual method is perceived as easier to use than textual one BUT the difference is not statistically significance.	*Threat-based visual method is perceived as easier to use than textual one.*
*RQ*2.2.2 PU	PU is higher for threat-based methods BUT the difference is not statistically significance.	Visual method is perceived as more useful than the textual one BUT the difference is not statistically significant.	*Visual method is perceived as more useful than the textual one.*
*RQ*2.3 ITU	Not tested.	*Participants intend to use the visual method more than the textual one.*	*Same as in Experiment 2.2.*

ties and we used Kendall's tau rank correlation coefficient to compare participants' actual effectiveness, PEOU, PU and ITU.

According to MEM there is a correlation between actual effectiveness and PU. To verify this we tested correlation between overall PU with the numbers and quality of identified threats and security controls for each of two methods and both methods in general. The results of the Kendall's tests show no statistically significant correlation between these variables, with a small exception when we consider the correlation between the numbers of security controls identified with correspondingly textual method and its overall PU ($p = 0.012, \tau = -0.25$) and visual method and its overall PU ($p = 0.025, \tau = 0.23$). However, it is not enough to conclude that actual effectiveness correlate with PU. The results of the Kendall's tests in Experiment 2.3 also revealed no correlation between the number and quality of threats and security controls with perception variables. Therefore, we cannot support corresponding MEM's claim. In contrast, correlations between PU, PEOU and ITU are statistically significant according to the results of the Kendall's tests both in the second and third experiments. Thus, our experiments supports MEM's claim in this respect.

2.8 Discussion and Implications

In this section we discuss the results of our validation of MEM and the main drivers behind MEM's constructs that we derived from the results of qualitative analysis. Table 2.14 also presents the main findings of both experiment regarding each research question.

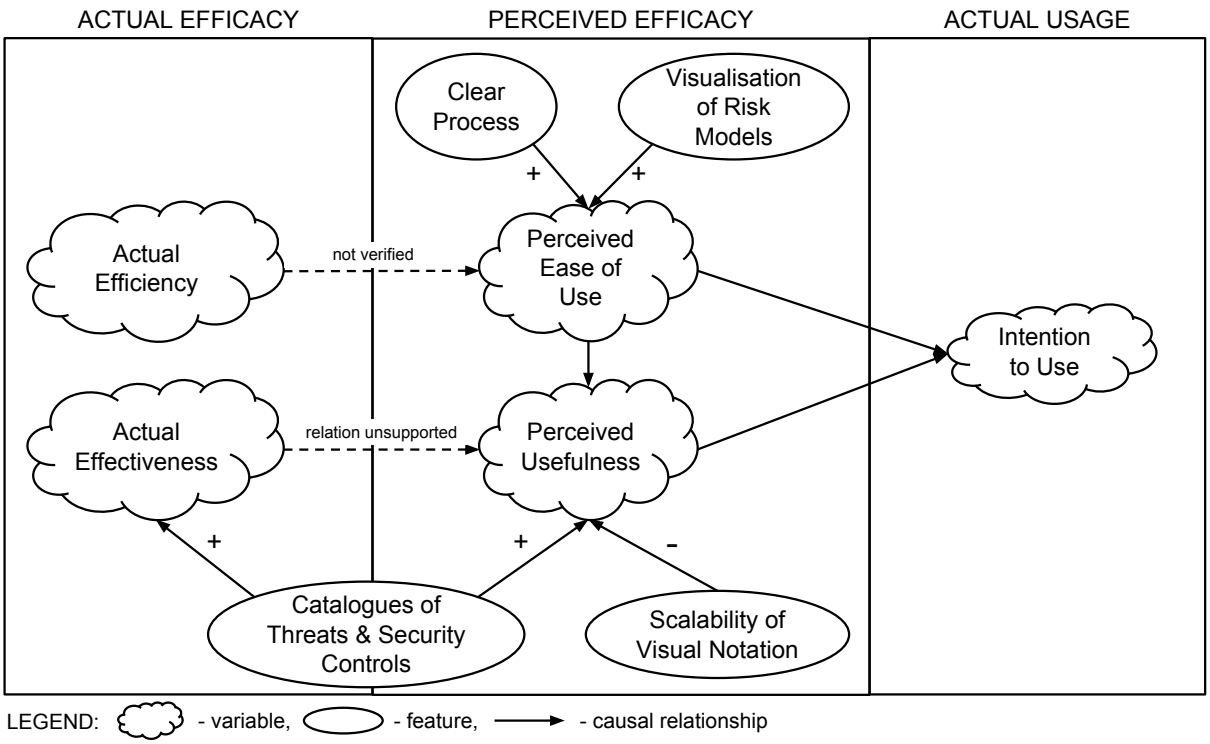


Figure 2.10: Refined Model

We have investigated the main drivers behind PU and PEOU of the methods, a part that is missing in MEM. The analysis results of focus groups and individual interviews, post-it notes, and open questions allow us to identify the main aspects impacting methods' PU and PEOU. These aspects are presented in Figure 2.10.

The main driver for methods' PEOU is presence in a method of *clear process* supporting main steps of SRA (identification of assets, threats and security controls). Also the importance of this aspect is confirmed by the results of correlation analysis of control question about process (see Q13 results in Table 6.11 on page 107 and question's statement in Table 6.17 on page 117) in Experiment 2.2 with overall PEOU of both methods (Kendall's test returned $p = 0.02, \tau = 0.24$). Based on the results of all three experiments we can conclude that availability of *visual summary* is reported as an aspect that positively impacts methods' PEOU. However, if visual summary *does not scale well*, it can harm methods' PU. If method *helps in identification of threats and security controls* than it can increase PU. For drivers related to help in identification of threats and security controls we have additional evidence from Experiment 2.2 based on the results of correlation analysis of control questions about help in brainstorming on threats and security controls (Q16 and Q17). The results support the fact that these drivers positively impacts methods' PU. We can also suppose that *help in identification of threats and security controls* can be increased with availability of *catalogs of threats and security controls* as

suggested by answers to post-task question $Q14$ in Experiment 2.2 and questions $Q2$ and $Q3$ in Experiment 2.3.

One of the main implications both for industrial practitioners and researchers that comes from the refined theoretical model is that there is a number of SRA specific features that should be taken into account for comparison or selection of methods: a) the presence of clear process supporting SRA steps, b) availability of visual summary and c) catalogues of threats and security controls. However, both methods' designers and users should be aware about scalability issues with visual representation that may appear in case of large systems modeling. A possible solution may be a tool supporting SRA steps and work with large models that decrease the effort required to model large systems.

2.9 Threats to Validity

We discuss the four main types of threats to validity [128] in what follows.

Conclusion Validity is concerned with issues that affect the ability to draw the correct conclusion about relations between the treatment and the outcome of the experiment.

– *Heterogeneity of participants.* If groups in a sample are too heterogeneous, the variation due to individual differences may be larger than due to treatment. We have reduced this threat by running experiments with groups which participants had similar knowledge and background. For Exp. 2.1 we had groups composed of at least one professional and one MSc student, while in Exp. 2.2 and 2.3 we had MSc students only.
– *Low Statistical Power* An important threat to validity is related to the sample size that must be big enough to come to correct conclusions. Since our sample in Experiment 2.1 is between 5 and 20 participants we used the Kruskal-Wallis test [53] and the Mann-Whitney U test [84]. For Experiments 2.2 and 2.3 we conducted a post-hoc power analysis with G*Power 3 tool [2] for participants from good groups. In Experiment 2.2 for the Wilcoxon signed-rank test we obtained a power (1-β) equal to 0.86 setting as parameter the effect size ES=0.71, the total sample size $N = 24$, and $\alpha = 0.05$. For the ANOVA test, we have instead a power of 0.89 with 32 observations for each method and between variance least 16 observations are needed to have an effect size of 2 like in our experiment. We thus have enough observations to conclude that our results on methods' actual effectiveness, PEOU, PU, and ITU are correct.
  In Experiment 2.3 for the results of Wilcoxon test with good participants we obtained a low power ($PEOU = 0.54, PU = 0.31$, and $ITU = 0.3$, where $N = 20$), while for all participant we received the following powers: $PEOU = 0.84, PU = 0.35$, and $ITU = 0.34$ with $N = 56$. To obtain 0.8 power we would need to have at least 40 participants for PEOU up to 96 good participants for ITU.

[2] http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

Internal Validity is concerned with issues that may falsely indicate a causal relationship between the treatment and the outcome, although there is none.

– *Bias in data analysis.* To avoid bias in reports analysis, coding of participants' reports was conducted by two researchers independently. In addition, the quality of threats and security controls identified by each group was assessed by experts external to experiments. In Experiment 2.1 we had two experts due to the presence of two application scenario while in Experiment 2.2 we had only one application scenario and one expert. However, in Experiment 2.3 we asked two experts to evaluate participants' results to have at least two opinions about results quality.

Construct Validity.

– *Experimenter expectancies.* The main threat to construct validity in our experiment is the design of research instruments: interviews and questionnaires. In Experiment 2.1 we measured participants' overall perception of methods. In Experiment 2.2 our questionnaire was designed following TAM with questions for each independent variable we wanted to measure: PEOU, PU and ITU. The interview guide included questions concerning $RQ2.2$ and methods' advantages and disadvantages. Several researchers have checked questions included in the interview guide and in questionnaires. Therefore, we believe that our research instruments measure what we want to measure. Moreover, to reduce this threat we have gathered data using other data sources (audio files, post-it notes, open questions and participants' reports) and performed data triangulation.
– *Hypothesis guessing.* Participants did not know which hypotheses were stated, and were not involved in any discussion on advantages and disadvantages of other methods, thus they were not able to guess what the expected results were.

Other threats to construct validity are considered small.

External Validity concerns the ability to generalize experiment results beyond experimental settings. External validity is thus affected by the objects and the participants chosen to conduct experiments.

– *Use of students instead of professionals.* Using students rather than professionals as participants is known as a major threat to external validity. In Experiment 2.1 we mitigated this threat by involving both MSc students and professionals that were working in groups. In Experiment 2.2 we mitigated this threat by using MSc students enrolled in a Security Engineering course. This allowed us to rely on students with the required expertise in security and to ensure that they had the same level of knowledge.
– *Realism of the application scenario and tasks.* We reduce the threat to external validity by making experimental environment as realistic as possible. In fact, as object of our experiment we have chosen two real industrial application scenarios proposed by Atos Research (Smart Grid) and Siemens (e-Health).

2.10 Conclusions

The chapter presented an evaluation framework to compare different SRA methods, a pilot study to test and refine the framework, and the results of three empirical studies conducted a) to compare three classes of academic methods to identify threats and security measures: threat-based methods (CORAS, SREP, LINDDUN,), goal-based methods (SECURE TROPOS), and problem-based methods (SECURITY ARGUMENTATION); b) to compare two types of threat-based methods: visual method (CORAS) and textual method (SREP and EUROCONTROL SecRAM). We compare methods with respect to actual effectiveness, overall perception, perceived ease of use, perceived usefulness and intention to use. Experiment 2.1 involved MSc students in Computer Science and security audit professionals who have applied different classes of methods to analyze security issues of industrial application scenarios. Experiments 2.2 and 2.3 were conducted with MSc students in Computer Science. They have applied visual and textual threat-based methods to conduct SRA of an industrial application scenario.

Experiment 2.1 shows that threat-based methods have higher overall perception and perceived ease of use than goal-based and problem-based methods. This could be due to the fact that these methods have a clearly defined process to identify threats and security controls and use a graphical notation to present results. These findings are confirmed by the results of Experiment 2.2. The first experiment has also shown that there is no difference in perceived usefulness of different classes of methods.

In Experiment 2.2 we found out that the visual method is better in identification of threats than the textual one. Also participants of Experiment 2.2 were intending to use the visual method more than the textual one. In contrast, Experiment 2.3 failed to reveal any difference between textual and visual methods with respect to their actual effectiveness. However, the results showed that participants reported higher preference for visual methods over the textual ones with respect to PEOU, PU and ITU.

Chapter 3

The Role of Catalogues of Threats and Security Controls in Leveraging Security Knowledge

In Chapter 2 we presented a theoretical model that hypotheses different characteristics of SRA methods that determine methods' actual efficacy and perceived efficacy. One of the aspects proposed by the model in Chapter 2 is the usage of catalogues of threats and security controls. This chapter aims to investigate the role of catalogues in an SRA and proposes a theory to explain how different catalogues' features contribute into an effective risk assessment process for novices and experts in either domain or security knowledge.

3.1 Introduction

SRA is a key step in the design of critical systems. But IS architects often lack the necessary security knowledge to identify all appropriate security risks. Even experts can forget to treat risks that might be relevant for a system. To alleviate this issue, industrial SRA methods and standards come with catalogues of threats and security controls. Essentially, catalogues are a form of knowledge reuse [111] created at community level [63] and made available to individuals. Security catalogues can be divided into *domain-general* and *domain-specific* catalogues. Table 3.1 presents some examples of these two categories of catalogues.

Table 3.1: Examples of the Catalogues of Security Threats and Controls

Type	**Catalogues**
Domain-general catalogues	BSI IT-Grundschutz, ISO/IEC 27002 and 27005, NIST 800-53
Domain-specific catalogues	PCI DSS for banking information security, EATM for security and safety in Air Traffic Management, OWASP for web application security, CSA Cloud Control Matrix for cloud security

The purpose of this chapter is to investigate how security analysts with different levels of expertise (novices, domain experts, security experts) can benefit from knowledge reuse in an SRA, and how effective knowledge reuse is. The expectations are that catalogue should reduce errors for security experts and should enable domain experts (as opposed to security expert) to perform a prima-facie SRA.

This chapter proposes a theory to explain how different features of catalogues contribute into an effective SRA process for novices and experts in either domain or security. We built the theory by using a grounded theory from interviews of security experts. First, it focuses on two types of knowledge involved in an SRA: community knowledge (in a catalogue form) and personal knowledge of a security analyst. Further, it explains *a*) the core tasks essential to successfully perform an SRA at different levels of expertise and *b*) the features of the catalogue needed for these tasks. At the end, the theory models the relationships between catalogues' features and actual and perceived efficacy of SRA methods. We conducted two controlled experiments aiming to provide empirical support to our theory.

The quantitative analysis shows that domain experts that are not security experts can obtain almost the same quality results as experts in both domain and security working without catalogues. Regarding perceived efficacy, for students without domain expertise domain-specific catalogues were perceived to be useful than domain-general ones because they provides exhaustive set of threats and security controls specific to an application domain.

In addition, the qualitative analysis of focus group interviews shows that non-experts and security experts have a different perception of catalogues. Non-experts found catalogues useful as starting point to identify threats and controls but at the same time they were concerned about the difficulty in navigating catalogues because there were no link between threats and security controls. Security experts instead found catalogues mostly useful because they provide a common terminology to discuss about threats and controls and they can be used to check completeness of results.

This chapter proceeds as follows. Section 3.2 discusses background and presents the related works; Section 3.3 presents grounded theory construction for the theoretical model of the effects of catalogues features on an SRA that is proposed in Section 3.4. Section 3.5 presents the research method; Section 3.6 presents the motivation of domain selection and describes the setting of the study, whose findings are presented in Sections 3.7 and 3.8. Threats to validity to our study are discussed in Section 3.9. The findings and implications are discussed in Section 3.10 and conclusion are presented in Section 3.11.

Table 3.2: Security Risk Assessment and Management Steps

Steps 2 through 8 are repeated until the recommended controls are sufficient to reduce the level of risk to the IT system and its data to an acceptable level.

Steps	Step description (from ISO 31000 standard)
Step 1: System Characterization	In this step, the boundaries of the IT system are identified, along with the resources and the information that constitute the system.
Step 2: Threat Identification	The goal of this step is to identify potential threat-sources and compile a threat statement listing potential threat-sources that are applicable to the IT system being evaluated.
Step 3: Vulnerability Identification	The goal of this step is to develop a list of system vulnerabilities (flaws or weaknesses) that could be exploited by the potential threat-sources.
Step 4: Control Analysis	The goal of this step is to analyze the controls that have been implemented, or are planned for implementation, by the organization to minimize or eliminate the likelihood (or probability) of a threat's exercising a system vulnerability.
Step 5: Likelihood Determination	The goal of this step is to derive an overall likelihood rating that indicates the probability that a potential vulnerability may be exercised within the construct of associated threat environment.
Step 6: Impact Analysis	The goal of this step is to determine the adverse impact resulting from a successful threat exercise of a vulnerability.
Step 7: Risk Determination	The purpose of this step is to assess the level of risk to the IT system. The determination of risk for a particular threat/vulnerability pair can be expressed as a function of threat-source likelihood, its impact and the adequacy of planned or existing security controls for reducing or eliminating risk.
Step 8: Control Recommendations	During this step, controls that could mitigate or eliminate the identified risks, as appropriate to the organization's operations, are provided.
Step 9: Results Documentation	Once the risk assessment has been completed (threat-sources and vulnerabilities identified, risks assessed, and recommended controls provided), the results should be documented in an official report or briefing that helps senior management, mission owners, to make decisions on policy, procedures, budget, management and system operations.

3.2 Background

3.2.1 Security Risk Assessment and Knowledge Reuse

SRA is a complex problem solving process. Table 3.2 describes the steps of a typical security risk assessment and management process based on the NIST 800-30 standard. The steps of the SRA process are usually supported by a security catalogues which are a form to encode expert knowledge that can be reused. However, the knowledge reuse practice is not well investigated in security [111]. In contrast, the importance of knowledge and its managements and communication is well understood in IS. The survey by [105] investigated knowledge management state-of-the-art in IS through a literature analysis of 94 knowledge management papers published between 1990 and 2000 in six IS journals. Overall, the authors concluded that IS research tends to *"adopt an optimistic view of the role of knowledge in organizations."* Later [106] extended the previous work and proposed a theoretical framework where knowledge can be considered as an asset that can be owned and transferred, and the role of knowledge is to progress individuals and organizations. [30] also argued that knowledge is a fundamental component for organizational processes, and organization structure should be designed to facilitate knowledge communication between workers. This idea is also supported in Software Engineering (SE) by [92] and by [88] who emphasized the importance of knowledge sharing practice.

Knowledge can be divided into *personal* and *community* knowledge [127]. Personal knowledge is tacit knowledge that people create by themselves or learn from their own

experience. Based on personal knowledge people make decision in their future projects. If people lack necessary knowledge they turn to community knowledge, which is "personal knowledge" shared between members, for example, in a documented form (catalogue being just one of such form). Indeed, a theory of knowledge reuse by [63] suggests that the work of experts can be facilitated by providing knowledge about proven solutions or best practices to problems in a new context. Knowledge reuse can also mitigate lack of expertise for novices or make easier work of professionals because they do not need to solve a problem again. For example, a catalogue of Non-Functional Requirements (NFR) was proposed as a part of NFR Framework to help developers to satisfy most common NFRs from SE practice [7]. The role of the knowledge in software security engineering is well described by [3], *"software security practitioners place a premium on knowledge and experience"*, who also discussed different types of *knowledge catalogues* in security, namely principles, guidelines, rules, attack patterns, historical risks, vulnerabilities and exploits. The authors suggested that security catalogues dissemination will help to refine and validate this knowledge and may be move the field toward standardization. [111] showed that threats and security requirements are the most reusable elements due to their importance for an SRA process. Usually these elements are presented in form of security patterns [98, 107] which can be also organized in catalogues.

3.2.2 Empirical Studies on Knowledge Reuse

The importance of reusable knowledge in its various forms is being advocated by academia and industry but very few studies have empirically investigated its effectiveness. [25] conducted empirical study with 69 software development teams which revealed that teams performance is strongly related to a knowledge communication practice adopted in teams.

There is no agreement in the literature whether external community knowledge (as captured by catalogues) is always effective in practice. For example in requirements analysis, the use of structured knowledge led to better coverage and completeness of gathered requirements [67]. However, the use of catalogue of NFRs needed to be coupled to a systematic method to result in significantly better performance in NFR elicitation than using only a catalogue or only a method [12]. In 1994 the "Gang of Four" published a book describing design patterns, solutions to common problems in software design [28] which became a bestseller in the SE community. Unfortunately, [134] showed that the Gang of Four patterns have limited usability and do not help novices to learn about design. Similarly, [133] was not able to demonstrate that the usage of security patterns improves neither the productivity of software designers nor design security. Business process improvements patterns were proposed to support users in application of improvements on business processes [24]. However, a combination of routing patterns and decision guidance for business process models creation was found to be time consuming due to increase of

efforts on the evaluation of different alternatives and decision making [129]. To the best of our knowledge this topic has not yet been investigated for SRA.

3.3 Qualitative Theory Construction for Security Risk Assessment Activities

In this study we used focus group interviews to collect data and grounded theory approach proposed by [32] to construct a theory based on collected data [36]. Figure 3.1 summarizes the approach we used. It follows the principle of emergence [120], according to which data gain their relevance in the analysis through a systematic generation and iterative conceptualization of codes, concepts and categories. Data are analyzed in constant comparison, broken into manageable pieces (codes) and compared for similarities and differences. Data that are similar in nature are grouped together under the same conceptual heading (category). Categories are developed in terms of their properties and dimensions and finally they provide the structure of the theory [120]. We approached this research by identifying a relevant problem currently experienced by security experts and practitioners in the field [3] and for which most empirical papers only focuses on comparing notations (e.g. graphical vs. textual). See for example [55, 58, 86, 118]. The major question from a knowledge management perspective is still unanswered: how codified community knowledge helps an SRA process?

As we can see from Table 3.2, SRA and management process involves activities like 1) identifying system boundaries, 2) finding/identifying/collecting necessary information, 3) validating and 4) presenting the results. To validate if these activities are really the core tasks for security risk assessment and management process, we conduct an empirical study with security professionals reported in the rest of the section.

3.3.1 Study Context

As a context for our study we selected SESAR ATM (Single European Sky ATM Research). SESAR is a public-private partnership which includes a total 70 organizations. SESAR coordinates and concentrates all EU R&D activities for future ATM research, including the development of its operational concepts (estimated at 2.1 billion euro).

This domain research is particularly interesting from a IS perspective as the technologies developed in the projects are mostly IS substituting existing physical systems. To illustrate this we provide some examples of IS in the scope of SESAR. For example, it includes the development of a fully remote control tower where physical out-of-the-window view is replaced by its digital version. Another projects, Airport Departure Data Entry Panel and Extended Arrival Management, aim at replacing manual assignment of landing and departure of flights by a fully automatic IS. SESAR Conflict Management and Au-

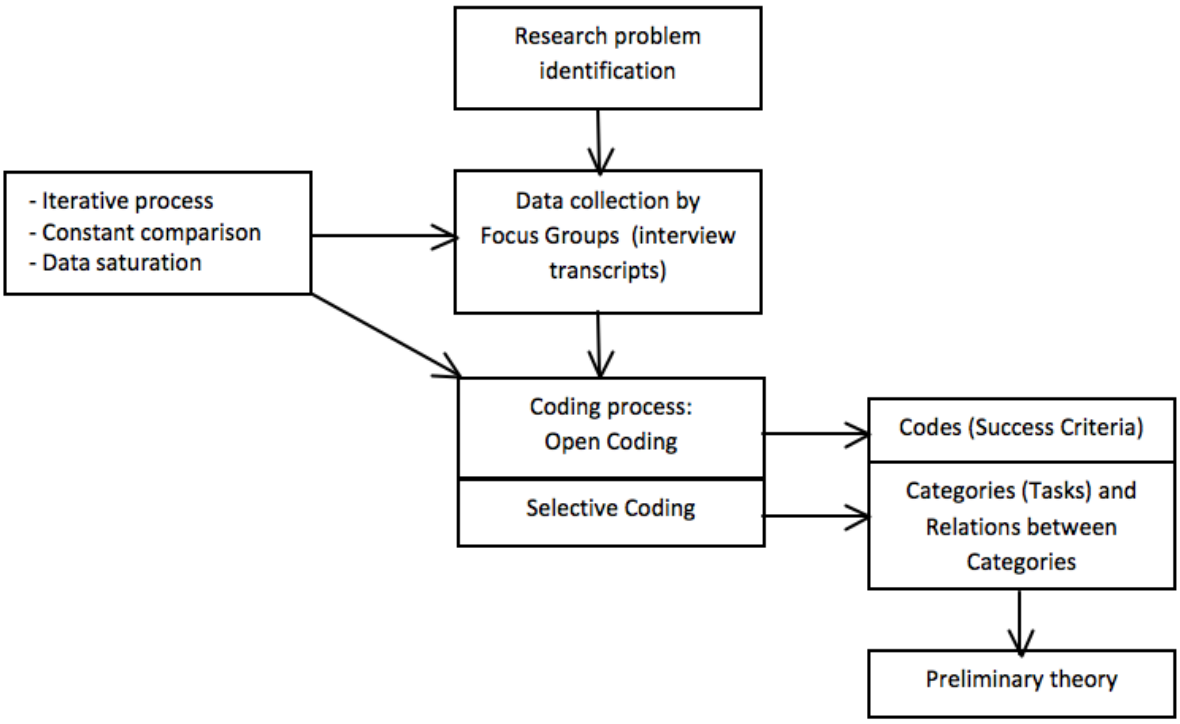


Figure 3.1: Research Approach

tomation project aims at significantly reducing controller work load through a substantial improvement of integrated automation support.

Our qualitative study involved Air Traffic Management (ATM) experts from SESAR Working Group. SESAR working groups include around 3,000 experts in Europe both in technological and organizational systems in ATM. We interacted with the experts in the Working Group in charge of the SRA for all developed solutions.

3.3.2 Data Collection

Various techniques exist for knowledge elicitation [42], but variation of structured or semi-structured interviews are most commonly involved in tasks analysis [113, Ch. 42]. The data analyzed in this research have been collected through a purposive sampling [38] from stakeholders attending the 6th Jamboree meeting of the SESAR Working Group in Brussels, 12th November 2013.

Table 3.3 presents demographic statistics about participants attending the meeting (total 20 experts). The participants were professionals with 17,5 years of working experience in average and in particular 7 years of experience in risk assessment. These participants can be defined as a small but representative selection of ATM and IS stakeholders carrying qualified opinions about and insights into SRA both physical and information security.

Table 3.3: Participants' Demographic Statistics

Variable	Scale	Mean	Distribution
Age	Years	42	5% were <30 years old; 58% were 30-45 years old; 37% were >45 years old
Gender	Sex	-	79% male; 21% female
Degree Education	-	-	6% High School; 17% BSc; 44% MSc; 17% MBA; 17% PhD
Professional Experience	Years	17,5	11% have <5 years; 37% have between 5 and 10 years; 53% have > 10 years
Professional Experience in Risk Assessment	Years	7	56% have <5 years; 28% have between 5 and 10 years; 17% have >10 years

We collected primary data from four parallel focus groups sessions [FG1, FG2, FG3 and FG4] lasted approximately 30 minutes. The participants were randomly assigned to the groups. In each group were 5 participants plus an individual moderator. The focus group interviews were audio recorded, then transcribed and coded with *Atlas.ti* software.

The focus groups were conducted in form of semi-structured interviews as they allow uncovering real interests perceived by participants rather than forcing a topic on them [132]. In order to extract type of knowledge, expertise and supporting artifacts needed to successfully perform the steps of an SRA method, the following areas of concern were used to facilitate keeping interviews in focus without biasing the responses from interviewees: *a*) aspects making an SRA method successful; *b*) weaknesses in SRA methods; *c*) factors influencing intention to use an SRA method; *d*) aspects making an SRA method easy to use; *e*) aspects making an SRA method effective; *f*) importance of compliance requirement when choosing an SRA method. The questions used to guide the discussion in the focus groups are reported in Table 7.3 on page 122 in the Appendix.

3.3.3 Data Analysis

Our analysis of qualitative data was inspired by the coding guidelines in [120, pp. 216-219]. The first phase of analysis (open coding) consists of collating the key point statements from each focus group transcript; a code summarizing the key points in some words is assigned to each key point statement. An example from our research is as follows:
Interview quotation: *"If you have different people do the risk assessment from different parts, you have a human factor [...]. Different experience, so this might somehow deviate the final result of the risk assessment."* (**Code**: Comparability of results).

In order to replicate the findings of another study we used the codes listed in Table 6.16 in Chapter 6; following the "iterative process" typical of coding method [94, Ch. 1] we finally identified 31 success criteria. Table 7.2 in the Appendix Chapter 7 reports the success criteria. A $*$ symbol marks codes different from the codes reported in Table 6.16[1]. We proceeded extracting from the success criteria some categories (selective coding) that we identified as the "tasks" needed to be performed by an expert during an SRA. "Finding Information", "Presenting/Sharing Information" and "Validating Information" have been

[1]We speculate that much difference may be due to the sample: here we used experts with 10+ years of experience, Table 6.16 presents the results of the qualitative study with MSc students

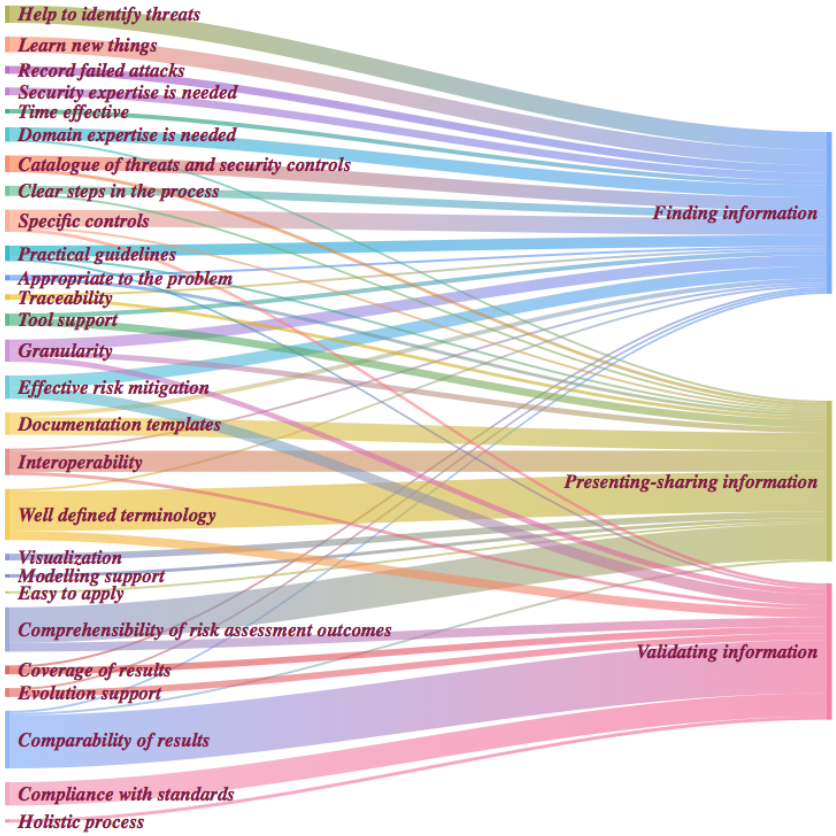


Figure 3.2: Co-occurrence between success criteria and tasks

selected as "tasks" emerged from the coding. The example below can explain how we moved from the quotation drawn from the interview to the code (success criteria) and lastly to the category (task):
Interview quotation: *"[The methodology] has to support the risk analysts in achieving the results they want, of course. So either identification of threats or estimation of likelihood or identification of security controls or whatever..."* (**Code**: Help to identifying threats; **Task**: Finding Information).
As a proxy for salience [36], in addition to presenting the relevance of each success criteria and task in terms of frequency in the interviews (Table 7.2 in the Appendix), we also calculated the frequency of their co-occurrence in the same statement. This is graphically shown in Figure 3.2.

3.3.4 Evidence from Interviews

Finding Information. Data analysis reveals that Finding Information is the core task of the whole risk assessment process: it is supported by the highest number of different success criteria identified by the participants. Its main task is to identify specific threats

and controls (FG1, 2, 3 and 4). In particular the methodology fulfills its own purpose when it allows to acquire some knowledge previously unknown (Learn new things): *"The best methodology leads to [...] [a] specific solution that is not covered by best practice"* (FG4).

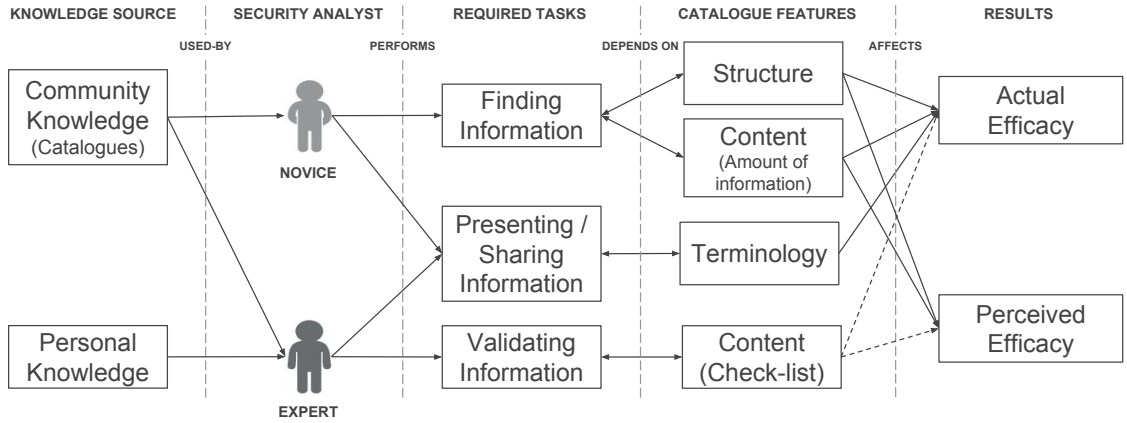
According to some experts a method is not stand-alone and it requires support also by other sources and tools: "The effectiveness is strongly linked with what are the tools that are around the methodology: the knowledge bases, the registers, the things that you build upon and you use to keep track of what you're looking for" (FG4). The availability of the Catalogue of threats and security controls is thus perceived to be important, as it can provide a good starting point in the analysis: *"The methodology should have a comprehensive threats catalogue so people may start from a base catalogue and then eventually add other threats"* (FG1). A catalogue can help in finding the right Granularity of the analysis: *"It will help you to identify the level of detail required to perform the risk assessment"* (FG3).

Among the other tools improving the methodology there are the Practical guidelines and the Documentation template: "It's not enough just to say: 'Okay, in this step the goal is this and that'. You need really to know exactly how to do it, and you should also have guidance on what kind of information you are supposed to gather during this step and how to document it, whether it's in a table or in a graphical format" (FG1).

Presenting/Sharing Information. The analysis of focus group discussion shows that when presenting the results, the most relevant aspect considered by interviewees is to have a well-defined terminology as this enhances the interoperability among experts and stakeholders and it is so important that: *"There is no sense if you have a super method, if the results cannot be exchanged [...]. You have to share the situation"* (FG2). The need for a common naming scheme is perceived as even more critical when interacting with, for example, customers, stakeholders, and regulators. This aspect is also related to the comprehensibility of method outcomes: " [A good methodology] allows me to explain to the person that's got to pay for the controls, what they need" (FG3). The existing gap between experts and customers at the comprehension and communication level is summarized in the need to provide to stakeholders what is called "the big picture" (FG2) and to "visualize" it (FG3). In this phase, documentation templates and catalogues of security controls as a baseline can support the presentation of the analysis results.

Validating Information. Participants involved in the discussion highlighted the importance of being compliant with standards: "[We shall] address governmental security needs and address the Austrian or French needs and so on" (FG4). Moreover, the results produced by the methodology implementation need also to be repeatable and comparable in order to be verifiable. This is perceived as important mainly in relation to repeatability when the context has changed to avoid the case that differences in the expertise of security risk assessment participants might affect final results of the risk assessment (FG2). This concern can also be addressed by using a well-defined terminology in this final phase.

Figure 3.3: A Theoretical Model for Catalogues as Knowledge Sources



Finally, the level of coverage of results provided by the methodology implementation is also important, as it should ensure a complete and exhaustive comprehensiveness of threats and security controls, as well as a coverage of the full life cycle of the system under threat.

3.4 A Theoretical Model for Catalogue Effectiveness

We summarize the findings of the previous section with a theoretical model of how users' expertise, identified tasks, and catalogues combine to produce an SRA. Figure 3.3 illustrates the key elements of the model and their relationships. The three tasks for SRA identified above are at its core: (a) *Finding information*, which implies identification of assets, threats and security controls, (b) *Validating information*, which means checking if the results produced by the analyst are complete and comply with standards, and (c) *Presenting/sharing information*, namely documenting results to other stakeholders using a terminology appropriate to the domain.

Novices are the main consumers of community knowledge. They use catalogues mostly to find information and adopt the appropriate language to present results. These activities are impacted by the following features:

- *Catalogue Structure.* If a catalogue does not have clear and logical structure it can affect novices' perceived efficacy and increase the effort needed to find the necessary information.

- *Catalogue Content (amount of information).* Novices can struggle with catalogues that are too big because they do not know how to start a risk assessment if they have too many options. Hence, the amount of information presented in a catalogue can affect both the actual and perceived efficacy of a security assessment.

- *Terminology.* A catalogue can support novices with standard terminology accepted in both the domain and the security field. Even users without solid background in domain or security can produce results understandable by experts (albeit the latter may disagree with the recommendations). The clear and uniform presentation of the results improves their actual efficacy.

Experts rely on their own knowledge as a source of information. They may benefit from the use of community knowledge to validate the correctness and coverage of their findings as well as to check the terminology used to present results to the particular customers. The following features play a major role:

- *Catalogue as a Checklist* may be useful when the user has experience ("personal knowledge") and needs to check that nothing important was overlooked. The completeness of the results has a positive effect on their actual and perceived efficacy.

- *Terminology.* Security experts often work across domains. They can use catalogues as a source of terms accepted in the domain, and thus ensure that results are understandable by relevant stakeholders, who are likely domain experts but not security experts, using an appropriate language. Thus, a domain-specific catalogue can make easier the work of experts in a new domain and can improve the actual efficacy of the results.

3.5 Experimental Validation

The goal of our empirical studies is to provide empirical support to the model depicted in Figure 3.3 and investigate whether different catalogues equally facilitate execution of an SRA. In particular, we want to assess whether the use of catalogues has an effect on actual and perceived efficacy of an SRA when used by people with no security expertise and comparing it with the effect of running the same assessment by security experts without catalogues.

We follow the protocol reported in [86] as amended by [55] to insert an expert evaluation of the artifacts produced by participants. Before execution of any activity, participants are administered a questionnaire to collect information on their background and previous knowledge of other risk assessment methods. Then the following steps are performed:

– *Training.* A scenario description is administered to participants by either individual reading or by an introductory presentation. Then, a frontal-training phase follows in which the expert in the method introduces the methodology to be used for the SRA through a step-by-step tutorial.
– *Application.* Participants apply the method to the scenario.
– *Evaluation.* External evaluators from industry assess the quality of threats and security controls identified by participants, providing marks and comments. These expert

evaluators are *neither* researchers nor their colleagues. They are independent industry experts contracted for the task.

After the application phase a post-task questionnaire is administered to participants to gather their perception of the method (and catalogues). Then they are involved into focus groups to discuss drawbacks and benefits of the method and catalogues they used. A list of questions is used to guide the discussion that is audio recorded for further analysis. The main positive and negative aspects reported in the focus groups then are reported on post-it notes by participants. The qualitative analysis attempted to cast light on catalogues' features affecting actual and perceived efficacy of SRA.

3.5.1 Treatment Groups

In the first experiment we only considered participants without significant domain expertise and therefore only divided them in two groups: the first group conducted an SRA with the support of a domain-specific catalogue (DOM CAT), the second group with the support of a domain-general (GEN CAT) one.

In the second experiment we only considered professionals and we created two groups as in the first experiments (DOM CAT, GEN CAT) and a third group which worked without catalogue (NO CAT). All participants in the NO CAT group had security knowledge, while most of the participants in the DOM CAT and GEN CAT groups had limited or no security knowledge.

3.5.2 Constructs and Measurements

The actual efficacy of a method can be measured in several ways. For example, [86] proposed to measure actual performance of the user by counting number of identified threats. This metric was also used by [51] and [97]. Similar metrics, e.g. number of identified failure modes or hazards, were adopted for safety analysis [115, 116, 118, 119]. Coverage is an alternative, more qualitative metric that is especially important for measuring effectiveness of methods for software testing [27,33]. In SRA, coverage is the type of threats identified [51,86] or the comparison of the proposed assessment against a baseline developed by an expert [78, 97].

In this work we measured *actual efficacy* as the quality of results produced by participants. Using the number of threats and security controls as a performance metric would be meaningless because there are lot of threats and security controls available in catalogues, and participants could include any of them in the analysis. Yet, they maybe irrelevant. This is also advocated by the experts who assessed results of our previous experiments: "*Threats are generic but understandable, although many threats are missing.*" and "*Very generic threats. Lack of understanding around the motivation of the threats.*" We therefore asked each expert to rate the overall quality of results on a 1-5 scale as fol-

lows: *Bad* (1), when it was not clear which are the final threats or security controls for the scenario; *Poor* (2), when threats/security controls were not specific for the scenario; *Fair* (3), when *some* of them were related to the scenario; *Good* (4), threats/security controls were specific for the scenario; and *Excellent* (5), when the threats were significant for the scenario and security controls propose an effective solution for the scenario. Figure 7.1 in the appendix reports the quality assessment guidelines agreed with the experts.

To assess *perceived efficacy* we used both the quantitative and qualitative approach. We first asked participants to fill in a post-task questionnaire. The questionnaire contains 10-20 questions about different constructs specific to perceived usefulness (PU) and perceived ease of use (PEOU) variables [14]. This approach was also applied to measure perceived efficacy of security and safety methods in numerous studies like [50, 79, 86, 118, 131]. Questions were formulated as opposite statements with answers on a 5-point Likert scale. Table 7.4 in the appendix reports the post-task questionnaire.

To validate the proposed theoretical model we also investigated transcripts of interviews on the basis of the set of codes already discussed in Section 3.3.

3.5.3 Data Analysis

In our study we are interested to prove the equivalence of different types of catalogues over expertise. Therefore, we use **equivalence testing** – TOST, which was proposed by Schuirmann [104] and is widely used in pharmacological and food sciences to answer the question whether two treatments are equivalent within a particular range δ [26, 75]. We summarize the key aspects of TOST as it is not well known in SE and refer to the review paper by Meyners [75] for details. The problem of the equivalence test can be formulated as follows:

$$H_0 : |\mu_A - \mu_B| > \delta \qquad \text{vs} \qquad H_a : |\mu_A - \mu_B| \leq \delta. \tag{3.1}$$

where μ_A and μ_B are means of methods A and B, and δ corresponds to the range within which we consider two methods to be equivalent.

Such question can be tested as a combination of *two* tests, as:

$$\begin{array}{lll} H_{01} : \mu_A < \mu_B - \delta & \text{or} & H_{02} : \mu_A > \mu_B + \delta \\ H_{a1} : \mu_A \geq \mu_B - \delta & \text{and} & H_{a2} : \mu_A \leq \mu_B + \delta, \end{array} \tag{3.2}$$

The p-value is then the maximum among p-values of the two tests (see [75] for an explanation on why it is not necessary to perform a Bonferroni-Holms correction). The underlaying statistical test for each of these two alternative hypothesis can then be any difference tests (eg. t-test, Wilcoxon, Mann-Whitney etc.) as appropriate to the underlying data.

Table 3.4: Descriptive Statistics of the Sample

Function	Min	Max	Mean	Median	σ
Mean of samples	1.48	4.05	2.92	2.96	0.68
σ of samples	0.35	1.28	0.86	0.835	0.25

In several cases for treatment equivalence it is preferrable to test for equivalence within a percentage value and namely test whether

$$H_{a1} : \rho_1 \cdot \mu_B \leq \mu_A \quad \text{and} \quad H_{a2} : \mu_A \leq \rho_2 \cdot \mu_B \tag{3.3}$$

However, when the percentage is applied to values on an ordinal scale it may harm equivalence analysis because the percentage of mean values ≤ 2 is significantly less than the percentage of mean values ≥ 4 on 1-5 scale. It means that samples with bigger mean values have higher chance to be found equivalent while the samples with smaller mean values are likely to be found non-equivalent. To eliminate this dependence on the mean values we decided to use an absolute value for δ.

To define δ we relied on an empirical approach and calculated δ that corresponds to a σ_p pooled across the samples reported in the literature. In our case we are looking for σ_p that estimates standard deviation of the variables on a 5-item Likert scale in experiments with people. To collect the sample of Information Systems studies we used Google Scholar search service, as it allows to search in the text of the papers, and the following criteria:

- Publication year: between 2010 and 2016.
- Journals: "MIS Quarterly" (MISQ), "INFORMS Information Systems Research" (ISR), and "INFORMS Management Science" (ManSci).
- Search terms: ("5-point scale" OR "Likert scale") AND ("standard deviation" OR "stdev").

The results of literature search are reported in Section 7.1 in Appendix. From the identified papers we extracted descriptive statistics of ordinal variables for 36 samples. Table 3.4 reports descriptive statistics of variables means and standard deviations across collected samples.

To calculate pooled σ we used the following formula:

$$\sigma_p = \sqrt{\frac{\sum_{i=1}^{k}(N_i - 1) \cdot \sigma_i^2}{\sum_{i=1}^{k}(N_i - 1)}}, \tag{3.4}$$

where N_i is the sample size and σ_i is the variance of sample i. Using (3.4) on collected dataset of 36 samples we obtained σ_p equals to 0.7. This value we adopted as the δ for equivalence test. The individual test chosen for the comparison is the Mann-Whitney rank sum test as we have independent samples.

3.6 Experimental Settings

3.6.1 Domain

The application scenario was chosen among one of the ATM new operational scenarios that have already been assessed by SESAR with SecRAM method: the Remotely Operated Tower (ROT). The Remote and Virtual Tower, is a new operational concept proposed by SESAR[2]. The main change with respect to current operations is that control tower operators will no longer be located at the aerodrome. The visual surveillance by the air traffic controller will be replaced by a virtual reproduction of the out-of-the-window view, by using visual information capture and/or other sensors such as cameras with a 360-degree view and overlaid with information from additional sources such as surface movement radar, surveillance radar, etc. LFV and Saab in Sweden did the first implementation of the ROT[3]

As apparent from the description, the ROT concept is a complex cyber-physical information system encompassing both by cyber-security issues (e.g. data confidentiality, integrity and availability of sensor data) as well as physical security issues, like on-site protection of the remotely located cameras and sensors. We think it is a good representation of security challenges faced by modern companies.

3.6.2 Method

We selected the SESAR ATM Security Risk Assessment Method (SecRAM)[4] as SRA method to be applied by participants for four main reasons: a) it is a method used in the ATM domain to conduct SRA of operation concepts, and its steps are very close to steps of other risk assessment standards, e.g. NIST 800-30; b) SESAR has conducted an SRA of the ROT operational concept with SecRAM that can be used to benchmark participants' results: c) the application of SecRAM is supported by the use of catalogues of threats and security controls; d) a SecRAM expert was available to train our participants; and e) the method was deliberately designed to be easy to understand by personnel with little expertise and background in security and risk management.

3.6.3 Catalogues

SecRAM supports personnel with catalogues of threats and security controls specific for the ATM domain (DOM CAT) developed by EUROCONTROL, the European Organisation for the Safety of Air Navigation. Our domain-specific catalogues have clear and

[2]SESAR Project P12.04.07: Single Remote Tower Technical Specification Remotely Operated Tower Multiple Controlled Airports with Integrated Working Position

[3]"LFV first in the world to have an operating license for remote towers" (`http://news.cision.com/lfv/r/lfv-first-in-the-world-to-have-an-operating-licence-for-remote-towers,c9672916`).

[4]SESAR Deliverable WP16.02.03: ATM Security Risk Assessment Methodology

Table 3.5: Catalogues' Main Characteristics

	DOM CAT	**GEN CAT**
Developer	EUROCONTROL	BSI
Threats (num. and types)	32 (Physical, Information and Procedural)	621 (Basic threats, Force Majeure, Organizational Shortcomings, Human Error, Technical Failure and Deliberate Acts)
Sec. controls (num. and types)	51 (Pre- and Post- controls)	1444 (Infrastructure, Organization, Personnel, Hardware and software, Communication and Contingency planning)
Link between threats and sec. controls	Yes (two-way); as a part of threats or sec. controls description	Yes (two-way); in a separate section

simple structure (32 threats divided into three topics with links to security controls), reasonable size (155 pages), support users with ATM specific terminology, and covers main problems related to ATM and proposes effective controls for them. For general catalogues we selected the BSI IT-Grundschutz catalogues developed (GEN CAT) by the German Federal Office for Information Security. It is compatible with the ISO 2700x family of standards. Domain-general catalogues have complex structure (621 threats and 1444 security controls in 6 topics with links between threats and controls in a separate section), big size (˜2500 pages), supports users with common security terminology, and cover a wide range of IT security problems and solutions. The main characteristics of two catalogues are summarized in Table 3.5.

Since DOM CAT catalogues are confidential materials for EUROCONTROL, participants received only a paper version of the catalogues and had to sign a non-disclosure agreement. To avoid differences in the use of the two type of catalogues, we provided a paper version also of GEN CAT catalogues (but not the page with detailed implementation of controls), but participants were allowed to access online the full version of the GEN CAT.

3.6.4 Demographics

The first experiment was conducted in February 2014. The participants were 18 MSc students from different universities in Europe. The participants worked in groups of two. Nine groups were randomly assigned to the treatments: five groups applied SESAR SecRAM method to the ROT scenario using EUROCONTROL ATM catalogues (DOM CAT), while the other four groups used BSI IT-Grundschutz (GEN CAT).

Table 3.6 presents descriptive statistics about the participants. A significant share of participants (44%) reported a limited working experience (at least 3 years), some participants (22%) reported ≤ 2 years of workings experience, and the rest did not report any working experience. Some participants (28%) reported that they were involved in security/privacy initiatives; the rest did not report any similar experience. Our participants had limited expertise in safety and security regulations, while in security technologies they reported a general knowledge. Our participants also had no prior knowledge of the ATM domain.

Table 3.6: Participants' Demographic Statistics – Experiment 3.1 (Students)

Variable	Scale	Mean	Distribution
Age	Years	25.06	33% were 21-24 years old; 67% were 25-29 years old
Gender	Sex		56% male; 44% female
Education Length	Years	5.17	44% had <5 years; 6% had 5 years; 50% had >5 years
Work Experience	Years	2.90	33% had no experience; 22% had 1-2 years; 44% had 3-5 years
Experience in Security/Privacy Initiatives	Yes/No	-	28% involved; 72% not involved
Expertise in Safety Technology	1(Novice)-5(Expert)	1.83	44% novices; 28% beginners; 28% competent users
Expertise in Safety Regulation and Standards	—"—	1.56	61% novices; 22% beginners; 17% competent users
Expertise in Security Technology	—"—	2.28	17% novices; 50% beginners; 22% competent users; 11% proficient users
Expertise in Security Regulation and Standards	—"—	1.89	33% novices; 44% beginners; 22% competent users
Expertise in ATM	—"—	1.06	94% novices; 6% beginners

Table 3.7: Participants' Demographic Statistics – Experiment 3.2 (Professionals)

Variable	Scale	Mean	Distribution
Age	Years	33.1	20% were 25-29 years old; 53.3% were 30-39 years old; 20% were 40 and older
Gender	Sex		66.7% male; 33.3% female
Academic Degree			73.3% had MSc degree; 26.7% had PhD degree
Work Experience	Years	7.9	26.7% had $\geq$ 2 and <5 years; 46.7% had $\geq$ 5 and <10 years; 26.7% had $\geq$ 10 years
Experience in Risk Assessment	Years	0.67	Three participants had 2 years, 1.5 years and 0.25 years, respectively
Security/Privacy Experience	Yes/No	-	60% had experience; 40% had no experience

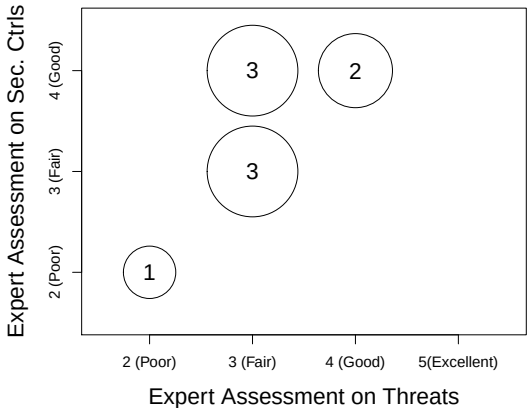
The second experiment was run in May 2014 at premises rented for the occasion and consisted of an empirical study with 15 professionals from several ATM Italian companies. As an incentive for professionals to participate, the activity was presented as a free training on SESAR SecRAM method for Risk Assessment by qualified experts[5]. The security trainings in ATM can be very expensive, e.g. a training on Aviation Cyber Security by the International Air Transport Association costs 2000 dollars. We divided participants into three groups and assigned to three different treatments. Then we asked them to apply individually the same method, namely SESAR SecRAM, with the support of DOM CAT, GEN CAT or without any catalogues (NOCAT).

Table 3.7 presents descriptive statistics about the participants. Most of the participants (73.4%) reported that they had at least 5 years of working experience, some participants (26.7%) reported from 2 to 5 years of workings experience. In addition, the majority of participants (60%) reported that they had security/privacy knowledge; the rest did not report any similar knowledge. Three out of sixteen participants reported from 3 months up to 2 years experience in SRA.

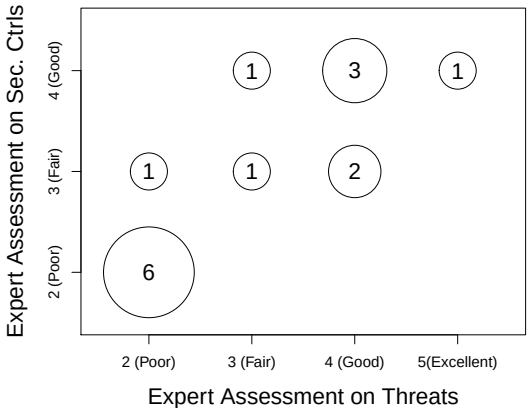
[5]Participants were aware that the results of their "exercises" would be used for research purposes.

Figure 3.4: Experts assessment of quality of threats and security controls.

(a) Experiment 3.1.



(b) Experiment 3.2.



3.7 Quantitative Results

The quantitative analysis investigates (1) whether different catalogues of threats and security controls have similar effect on the execution of an SRA and (2) whether the use of catalogues has an effect on the actual and perceived efficacy of SRA when used by people with no security expertise, and comparing that with the effect of running the same assessment by security experts without catalogues – essentially determining whether the upper chain of Figure 3.3 ("community knowledge + novice") can obtain the same results as the lower chain ("personal knowledge + expert"). We also investigate whether the features of the catalogue make a difference.

Quality Evaluation. Two ATM security experts independently assessed the quality of the results collected in the first experiment with students. They reported a similar assessment for each group. Only one group out of nine performed "poorly". Note that the expectations in terms of results of the assessment were higher for domain experts. Figure 3.4a illustrates the average of experts' evaluation for threats (reported on the x-axis) and security controls (on the y-axis) for the participants of the first experiment.

Figure 3.4b illustrates the average of experts' evaluation for threats (reported on x-axis) and security controls (on y-axis) for the second experiment with ATM professionals. Six participants out of fifteen performed poorly. In terms of the final assessment we observed that: *a)* the experts marked bad participants the same way; *b)* they consistently marked moderately good participants; and *c)* they had a different evaluation only for the threats of one participant and for the security controls of another participant out of 15 participants. The best results one of the experts commented with the following statement: "*Threats cover wide range including technical physical, social engineering and*

personnel issues. Controls demonstrate defense in depth and holistic approach. Excellent because the only one to have a threat relating to ATM (loss of aircraft separation)". Even when experts would slightly disagree on the overall mark, they would actually agree in the comment on the deficiencies of the evaluated work. For example, the first expert assessed security controls of one participant as Bad: "*For controls, Bad because simply not reasonable to accept risks where the impact is high (e.g. jamming radar)*", while the second expert put Fair for the security controls of the same participant: "*Some controls present but in most of cases risk was classified as tolerable and in consequence no PE or PO Controls were identified.*" Hence, for a quantitative study we can use the average of experts' votes.

Tables 3.8 to 3.10 report the descriptive statistics of the dependent variables for Experiments 3.1 and 3.2, the corresponding values of the statistical tests for equivalence with TOST over MW and difference with MW, and effect size d.

We present the main quantitative findings of the two experiments as follows:

We summarize the main quantitative findings of the two experiments as follows:

AE_1 *There is no difference in actual efficacy of catalogues for people without domain expertise.* This is supported by the first experiment in which both groups using domain-specific and domain-general catalogues delivered threats and security controls of similar quality. The comparison of DOM CAT vs GEN CAT has a $p = 0.056$ for the TOST on threats and $p = 0.087$ for controls. We only have significance at the 10% level. Domain experts with no security experience using catalogues identified threats and controls (2.5 as mean) of a slightly lower quality then security and domain experts without catalogues (2.8 as mean) but the results are not statistically significant (TOST $p = 0.15$ for threats and $p = 0.12$ for controls).

AE_2 *For domain experts, use of catalogues improves the quality of identified threats and security controls.* This is supported by the second experiment, in which domain and security experts used catalogues and delivered threats and security controls of better quality as people with domain and security expertise but without catalogues. The quality of results identified by the group with catalogues is better than for the group without catalogues and this is statistically significant at $p = 0.02$ for threats and $p = 0.03$ for controls with very large effect size.

PU *People without domain expertise think that domain- specific catalogues are more useful than domain-general ones.* This is supported by the first experiment, in which people who applied method with domain-specific catalogues reported higher PU of the method than participants who used domain- general catalogues. This is confirmed by a MW test that returned $p = 0.05$ for threats with large effect size according to Cohen's criteria, while TOST returned $p = 0.13$. For security controls results are inconclusive. The results of statistical tests showed that PU of two catalogues is equivalent at 10% significance level and PU of domain-specific catalogues is signif-

Table 3.8: Experiment 3.1 (Novices): Summary of Quantitative Results

For people with no domain experience there is a 10% significant equivalence between a specific or a general catalogue with respect to actual efficacy (AE) and perceived ease of use (PEOU); domain-specific catalogues have slightly better perceived usefulness (PU) than general catalogues.

Threats	DOM CAT			GEN CAT			Statistical Tests		Eff. size
	μ	med	σ	μ	med	σ	$TOST_{MW}$	MW	d
AE	3.33	3.33	0.63	3.08	3.00	0.17	0.056	0.24	0.51
PU	3.60	3.67	0.47	3.11	3.28	0.49	0.13	**0.05**	1.03
PEOU	3.49	3.69	0.82	3.67	3.69	0.70	0.07	0.81	-0.24

Sec. Ctrls	DOM CAT			GEN CAT			Statistical Tests		Eff. size
	μ	med	σ	μ	med	σ	$TOST_{MW}$	MW	d
AE	3.27	3.33	0.64	3.58	3.67	0.17	0.087	0.50	-0.64
PU	3.66	3.67	0.40	3.22	3.33	0.41	0.08	**0.03**	1.07
PEOU	3.52	3.62	0.74	3.69	3.69	0.64	0.08	0.74	-0.23

Table 3.9: Experiment 3.2 (Domain Experts): Results for Non-security Experts with Catalogues and Security Experts without Catalogues

For people with domain experience a catalogue improves the results of non security expert in comparison to security experts but not enough to make it equivalent in terms of selected δ. Regarding perceived efficacy both non- security and security domain experts reported PU and PEOU equivalent with 10% significance level.

Threats	No Sec. Expert CAT				Sec. Expert. NO CAT				Statistical Tests p-value		Effect Size
	N	μ	med	σ	N	μ	med	σ	$TOST_{MW}$	MW	d
AE	6	2.50	2.50	0.71	5	2.80	2.50	0.45	0.15	0.50	-0.50
PU	6	3.33	3.50	0.66	5	3.77	4.00	0.46	0.25	0.23	-0.75
PEOU	6	3.20	3.30	0.59	5	3.64	3.80	0.52	0.31	0.36	-0.78

Sec. Ctrls	No Sec. Expert CAT				Sec. Expert. NO CAT				Statistical Tests p-value		Effect size
	N	μ	med	σ	N	μ	med	σ	$TOST_{MW}$	MW	d
AE	6	2.50	2.50	0.45	5	2.80	3.00	0.57	0.12	0.40	-0.59
PU	6	3.31	3.43	0.61	5	3.77	3.71	0.41	0.26	0.21	-0.87
PEOU	6	3.00	2.90	0.55	5	3.64	3.80	0.52	0.58	0.07	-1.19

Table 3.10: Experiment 3.2 (Domain Experts): Results of Security Experts with Catalogues and without Catalogues

For people with domain and security experience a catalogue improves the results over security experts who conducted SRA without catalogues. This difference is statistically significant for AE of both threats and security controls with very large effect size. For PU of threats and controls and for PEOU of threats both groups reported results equivalent at 10% significance level. For PEOU of controls the results neither are equivalent nor different.

Threats	Sec. Expert CAT				Sec. Expert NO CAT				Statistical Test p-value		Effect size
	N	μ	med	σ	N	μ	med	σ	$TOST_{MW}$	MW	d
AE	4	4.12	4.00	0.63	5	2.80	2.50	0.45	0.95	**0.02**	2.49
PU	4	3.64	3.64	0.44	5	3.77	4.00	0.46	0.087	0.75	-0.28
PEOU	4	3.50	3.60	0.53	5	3.64	3.80	0.52	0.095	0.69	-0.27

Sec. Ctrls	Sec. Expert CAT				Sec. Expert NO CAT				Statistical Test p-value		Effect size
	N	μ	med	σ	N	μ	med	σ	$TOST_{MW}$	MW	d
AE	4	3.75	3.75	0.29	5	2.80	3.00	0.57	0.72	**0.03**	2.02
PU	4	3.68	3.50	0.41	5	3.77	3.71	0.41	0.095	0.69	-0.23
PEOU	4	3.30	3.40	0.66	5	3.64	3.80	0.52	0.19	0.61	-0.58

icantly better then PU of domain-general catalogues at $p = 0.03$. But this is not tested for domain experts as we have a small number of participants in the second experiment.

PEOU *There is no difference in PEOU of catalogues for people without domain experience.* This is supported by the first experiment, where people with domain-specific and with domain-general catalogues reported similar PEOU. This results is statistically significant at 10% level as TOST returned $p = 0.07$ for PEOU of threats and $p = 0.08$ for PEOU of controls. Due to small sample in the second experiment this results cannot be validated with domain experts.

The detailed results of risk assessment delivered by the participants are reported in Tables 7.5 and 7.6, and the detailed statistics on post-task questionnaire responses are reported in Tables 7.7 and 7.8.

3.8 Qualitative Results

The qualitative analysis clarifies how the relationships and catalogue features proposed in Section 3.4 are supported by the results of focus groups interviews with participants and post-it notes sessions summarizing each discussion within groups in the second experiment.

One issue identified in the analysis is the difference in opinions of security-experts and security-novices about their general perception towards the catalogue. Security-novices tend to express a positive judgment on the benefits of using catalogues. In contrast, security-experts tend to be more uncertain about actual advantages of catalogues. This could be explained by the fact that catalogue represents an essential support for users with no or little experience, as argued by one participant: "*The catalogue is really helpful if you do not have any background*".

Catalogue Structure. The analysis of interviews shows that the structure of a catalogue is a key aspect in the identification of threats and security controls. Thanks to its basic layout, clear tables, and its relatively short length, the domain-specific catalogues are generally perceived by participants as easier to browse and to read: "*I read only the titles* [namely the reference to the "Generic Threat" and the "Attack Threat"], *they were quite explanatory, therefore a very short consultation of the catalog allowed me to produce enough content*". This is particularly true in comparison with the domain-general catalogues, consisting of a long list of items, perceived as "*not user-friendly at a first read*" and "*difficult to navigate and master due to its length and structure*".

Another relevant aspect in the structure of catalogues is the presence of linking references between threats and security controls. According to some participants this feature makes the identification of controls an automatic mechanism: "*Once identified the threat, finding out controls was really a mechanical work*". Even more so for security-novices, traceability is perceived as a fundamental feature in catalogues structure. It provides a

one-directional link between the two objects of interest that makes mistakes difficult. In contrast, the domain-general catalogues do not provide this support and therefore the findings are affected: "*The identification of security controls was more difficult because you had to map them with the threats previously identified but there was no direct link in the catalogue. It was mainly due to a problem of usability of the catalogue*". Examples, present in the specific-domain catalogues, are also perceived as helpful for identification of threats and security controls.

Catalogue Content (amount of information). Even if a catalogue is meant for security-novices providing too many details and too much information may be counterproductive. Security-novices can feel overwhelmed and not able to find any threat or security control at all. This is particularly the case of the general-domain catalogues, judged as: "*Very difficult to consult for non-technical people*" given the high number of threats and controls proposed. An interesting statement in this regard, comes from a participant who was not assigned to any catalogue but had a chance to glance at the general-domain catalogues: "*I saw people near to me; they were not able to find out stuff in the catalogue, they kept on getting lost in the pages and eventually they came up always with the same two or three items*".

Catalogue Content (Check-list). Regarding the ability of catalogues to cover a variety of threats and controls, the opinions expressed by participants were quite varied: security experts claimed that the suggestions in both catalogues were very generic, rather than specific, precise and well-defined threats and controls: "[The catalogue provided a] *list of non-specific threats impacting the specific concept under investigation*" (from a domain-generic catalogue security-expert user). The same result comes from the domain-specific catalogues: "*I found the catalogue useful, but I noticed that many threats were repeated*". In contrast, security-novices were in general more satisfied by the use of catalogues. This is probably due to the fact that, without any experience any kind support is of great benefit and that participants could not really judge the quality of the catalogue itself.

The statements collected from security-experts suggest an additional aspect: "*The first step is to use your own experience and then to use the catalogue to cover generic aspects that could be forgotten*". For security-experts the catalogue is perceived as a check-list, as something that can be used after a brainstorming session where users work based on their own experience. In this way, the catalogue is supposed to validate the efficiency and coverage of the identified threats and security controls. For security-novices on the contrary, the catalogue represents: "*A good starting point for the evaluation of the threats and the controls.*".

Catalogue Terminology. One feature of the catalogue perceived as essential by every participant, irrespectively of the type of catalogue employed, is the fact that a catalogue by itself provides a common terminology for all users. As suggested by one participant, "*The catalogue could be seen as a useful tool, able to formalize the controls*

that have been formulated in an informal way, and to lead them back into a common nomenclature". "*The problem arises when we are in the same group and we use a different language*". The demand for a standard language caused by the need of sharing, discussing and presenting results by all stakeholders is an important feature of the risk assessment process. Unsurprisingly, this aspect is mostly perceived as important by participants who were not assigned to any catalogue.

In summary, participants with security knowledge cared more about the quality of threats and security controls that they could identify with the support of the catalogues. They expected more specific results from the support of the catalogue. Security-novices were not able to judge the quality of the identified threats and controls. Therefore, they were more concerned about catalogues' usability, as demonstrated by their observations on the traceability and navigability of the catalogues.

3.9 Threats to Validity

This section discusses threats to validity of the theoretical model and experimental results. The critical question in qualitative studies is the generalizability of the findings. We built our theoretical model using grounded theory analysis of the data collected in focus groups interviews with security experts. We expect that this model can be validated by other studies that maintain a settings similar to ours. For example, similar observations can be expected if the focus groups include other people that share same context (i.e. background, working experience, domain expertise) with those who participated in our focus groups.

In what follows we discuss threats to validity of experimental results. Domain-general catalogues are significantly larger than the domain-specific ones as they cover a wide variety of scenarios. We mitigated this threat to internal validity by making available domain-specific catalogues of relatively large size (155 pages) and by preparing an index of the general catalogues (~55 pages) that contained the list of available threats and security controls for ease of reference. Participants had also access to the full version of the domain-general catalogues (~2500 pages).

The main threat to conclusion validity is related to the *sample size* that must be big enough to come to correct conclusions. Our low number of participants ($N_{Exp1} = 18$ and $N_{Exp2} = 15$) makes it difficult to draw strong conclusions. [74] shows that it is possible to have statistically significant results also for small samples. To understand possible effect of participants' background on the results we collect information about participants' through demographics and background questionnaire at the beginning of the study. To mitigate possible previous knowledge about object of the study the participants were given a step-by-step tutorial on the SRA method and received a textual description of the application scenario.

Another threat to conclusion validity could be the number of security analyses con-

sidered as low quality by experts in the second experiment (6 out of 15). However, we think the level of quality illustrates the diversity of participants' knowledge and expertise. Therefore, we have different experiences of SRA. It could be a threat to validity if we had all risk assessments of the same quality.

The main threat to external validity is that both the risk assessment method and scenario were chosen within the ATM domain. However, the chosen risk assessment method is compliant with ISO 27005 standard that can be applied to different domains not just to the ATM. Therefore, this threat is fairly limited in our study. Another threat to external validity are the participants selected to conduct the experiment. We limit this threat by using professionals from ATM companies in addition to MSc students. Another threat to external validity is the *realism of experimental settings*. Our experiment significantly counters this threat in comparison to the literature [52,79,109,118] as we had the duration of two days rather than a couple of hours. This longer duration suggested by [55] allowed us to use a complex enough application scenario and thus to generalize our results to the real projects. In addition to the longer duration, we limited threats to conclusion validity because *a)* participants were trained by an expert in the method who usually trains professionals working in the ATM domain, and *b)* participants had two full days to apply the method to a new ATM operational concept.

3.10 Discussion and Implications

Our study shows that the use of catalogues can mitigate a lack of security expertise and provide a good starting point for the analyst. The results showed that in quantitative terms there is little difference between novices in actual efficacy of an SRA method when used with domain-specific or domain- general catalogues. A domain-specific catalogue was perceived to be more useful than a domain-general by novices with statistical significance. While professionals perceived domain-specific catalogue to be slightly more ease to use than a domain-general one but this result is not statistically significant. A more interesting result is that in the second experiment with ATM experts: there is a small difference in actual efficacy of a security method when used with catalogues by non-experts and without catalogues by security experts. Albeit only few groups achieved a high quality score in terms of identified threats and security controls. Also the security and domain experts identified threats and security controls of significantly higher quality using catalogues in comparison to security and domain experts performed SRA without catalogues. It shows that to limited extent security expertise can be codified and used by domain experts.

The qualitative analysis with ATM professionals, carried with focus group interviews and a post-it notes session, supported the relationships proposed by the theoretical model. Non-experts were mostly worried about the difficulty of navigating through catalogues while expert users found it mostly useful to get a common terminology and a checklist

that nothing was forgotten. Catalogues could provide support for discussion among the analysts because they provide a common language for analysts with different background. They could also be used to check the completeness and coverage of the results.

The main managerial implication that comes from the results of our study is that non-experts with catalogues can deliver results of a comparable quality to those produced by security experts. Thus, to facilitate the security analysts we should have a method that supports catalogues usage from the first steps. Usually, an SRA process requires to identify three main components: *1)* assets that should be protected, *2)* threats that can harm identified assets, and *3)* security controls that can mitigate identified threats. Catalogues can provide an ample source of knowledge for all three components. Analysts just need to limit scope to the assets which are relevant to the system and in this respect domain knowledge is all that is needed. Consequently, a set of preliminary threats and security controls can be identified by using catalogues. Thus, catalogues facilitate a prima facie SRA by domain expert. From a company's perspective domain experts are easier to find internally than security experts who are expensive to get[6].

However, such conclusion should not be stretched to present knowledge sourcing through catalogues as a complete solution. Indeed, our first experiment and our qualitative analysis showed that complete novices are not entirely better off. In particular, general, hard to search catalogues, which are the ones that novices in both domain and security knowledge are likely to download from the internet, does not seem to warrant a similar effectiveness. Finding effective source of knowledge requires more than simply collecting solutions or problems.

3.11 Conclusion

Security catalogues are an important part of the SRA process: *"as the [security] field evolves and establishes best practices, knowledge management can play a central role in encapsulating and spreading the emerging discipline more efficiently"* [3].

The aim of catalogues of threats and security controls is to put best security practices into a uniform format that can be re-used. In this chapter we have presented a theoretical model for the impact of codified knowledge (catalogues) on SRA process. We have investigated in both qualitative and quantitative terms the effect of using domain-specific catalogues versus domain-general catalogues, and have compared them with the effects of using the same method by security expert but without catalogues.

In summary, the study shows that with the use of catalogues a satisfactory number of threats and controls can be identified. If security expertise is expensive to get, a domain-specific catalogue is your second best bet.

[6]"Cybersecurity Professional Trends: A SANS Survey", SANS Institute, 2014. URL `https://www.sans.org/reading-room/whitepapers/analyst/cybersecurity-professional-trends-survey-34615` (Last accessed: March 2016).

Chapter 4

The Comprehensibility of Security Risk Modeling Approaches

This chapter aims to further investigate the aspects of the theoretical model presented in Figure 2.10 in Chapter 2 and answer the question: *" how comprehensible are different representation approaches for risk models?"*

Tabular and graphical representations are used to communicate the results of SRA for IT systems. However, there is no consensus on which type of representation better supports the comprehension of risks (such as the relationships between threats, vulnerabilities and security controls). Cognitive fit theory predicts that spatial relationship should be better captured by graphs.

4.1 Introduction

Security risk analysis plays a vital role in the software development life cycle because "it provides assurance that security concerns are identified and addressed as early as possible in the life cycle, yielding improved levels of attack resistance, tolerance and resilience" [70]. Risk analysis is usually performed by security experts but its results are consumed by 'normal' IT professionals (from managers to software architects and developers).

Presenting and communicating risk to all stakeholders is a key step to make sure risk analysis is not an empty exercise (e.g. it is an explicit step out of nine in the US NIST 800-30 standard process). This is particularly challenging as risk analysis tries to link a multitude of entities into a coherent picture: threats exploit vulnerabilities to attack assets and are blocked by security controls; attacks may happen with different likelihood and may have different levels of severity; one vulnerability may be present in several assets and an asset may be subject to several threats; security controls must address and reduce risks to acceptable levels in an optimal manner. Hence, the representation of security risk assessment results should be clear to all involved parties, from managers to rank-and-file

developers otherwise, they "[. . .] may find themselves lost in the process, misinterpreting result, and unable to be a productive member of the team." [59, p. 45]. A qualitative empirical study on the success criteria for security risk assessment with professionals with 17.5 years of work experience on average and in particular 7 years of experience in risk assessment highlighted communication as one the key features [57, Table 2].

Existing risk analysis methods and techniques use different notations to describe the result of risk analsys. Industry methods typically use a *tabular modeling notation* (eg. ISO 270001, NIST 800-30, SESAR SecRAM, SREP [71]) whereas academic based methods use *graphical modeling notations* (eg. SI^* [31], Secure Tropos [81], ISSRM [66], or CORAS [61]). Yet, there is limited empirical evidence whether one of the two risk modeling notation better supports the comprehension of security risks. Hence, this chapter aims to investigate the following research questions:

*RQ*4.1 *Which risk modeling notation, tabular or graphical, is more effective in extracting correct information about security risks?*

*RQ*4.2 *What is the effect of task complexity on participants' actual comprehension of information presented in risk models?*

To answer these research questions we have conducted two studies with 69 and 83 students. The first study consisted of three experiments: one performed at the University of Trento, Italy, and two performed at PUCRS, Porto Alegre, Brazil. In Trento, the experiment involved 35 graduate students; in Porto Alegre, the two experiments were run with 13 graduate and 21 undergraduate students. The second study included two experiments: one performed at the University of Calabria in Cosenza, Italy, the experiment involved 52 master graduates attending a professional post-master course in Cybersecurity, and the second one at the University of Trento with 51 master students attending a Security Engineering course.

We considered comprehension tasks of different complexity in line with Wood's theory of task complexity [130]. We selected scenarios from the healthcare and online banking domains, modeled the security risks of the scenario in the two modeling notations, and asked the participants to answer several questions of different level of complexity. By using the metrics of precision and recall on the answers provided by participants we compared the effect of the modeling notation and other potential factors (education, modeling or security experience, knowledge of the English language) on the comprehensibility of the risk models.

The remainder of the chapter is organized as follows. Section 4.2 discusses related work. Section 4.3 describes the study design. Section 4.4 reports the experiments realization. Section 4.5 presents the analysis results and Section 4.6 discusses their implications. Section 4.7 discusses the threats to validity of our study.

4.2 Related Work

Several studies have compared textual and visual notations: some studies have proposed cognitive theories to explain the differences between the two notations or to explain their relative strengths [76, 125]; other studies have compared different notations from a conceptual point of view [48, 95]. Several empirical studies have compared graphical and textual representations for requirements [108, 116, 118, 119] , software architectures [41], and business processes [87]. Studies that focus on comparing textual and visual notations for security risk models are less frequent [35, 43] or compared the effectiveness of tabular or graphical methodologies as whole [55, 58, 65] as opposed to the specific aspect of comprehensibility.

4.2.1 Empirical Comparisons of Software Modelling Notations

Among the works which reported empirical studies on the effectiveness of visual vs. textual notions focusing on the early stages of software development Hoisl et al. [44] compared three notations for defining scenario-based tests (a semi-structured natural-language notation, a diagrammatic notation, and a fully structured textual notation). The metrics considered accuracy and effort involved in understanding scenario-test definitions, and detection of the errors in the models under test. The results of the study showed that the participants who used the natural-language notation spent less time and completed the task with higher accuracy than the participants who used the other two notations. Participants also expressed higher preference for the natural-language notation. Based on the results of the ex-post questionnaire, the authors concluded that possible explanations of these results could be that (1) the diagrammatic notation has poor scalability and for complex scenarios it becomes hard to understand, and (2) fully structured notation needs specific preparation and additional materials in order to be understood.

Scanniello et al. [99] conducted four controlled experiments with students and professional to investigate the effect of UML analysis models on comprehensibility and modifiability of source-code. The participants were asked to complete tasks using both treatments (i.e., having source code and analysis models and having source code only) for two different systems to control learning effect. The results revealed no difference in understanding source code and ability to modify it with and without having UML analysis models. The authors explained the results by the fact that the provided UML models did not contain any details on the systems implementation, and therefore, not very helpful for understanding and modifying source code.

Sharafi et al. [108] assessed the effect of using graphical vs. textual representations on participants' efficiency in performing requirements comprehension tasks. They found no difference in accuracy of the answers given by participants who used the textual and the graphical notations but it took them considerably more time to perform the task with a

graphical notation than with textual one. Still, the participants preferred the graphical notation. Surprisingly, the participants spent significantly less time and less effort while working on the third model with both graphical and textual representations than with the other two models. The authors explained this finding as being due to the fact that the participants learned the graphical notation after performing the comprehension task which led to the improved results with the mixed model. Similarly, Abrahao et al. [1] assessed the effectiveness of dynamic modeling in requirements comprehension. The study included 5 controlled experiments with 112 participants with different levels of experience. The paper revealed that providing requirements specification together with dynamic models, namely sequence diagrams, significantly improves comprehension of software requirements in comparison to having just specification document.

Heijstek et al. [41] investigated the effectiveness of visual and textual artifacts in communicating software architecture design decisions to software developers. Their findings suggest that neither visual nor textual artifacts had a significant effect in that case. Ottensooser et al. [87] compared the understandability of textual notations (textual use cases) and graphical notations (BPM) for business process description. The results showed that all participants well understood the textual use cases, while the BPMN models were well understood only by students with good knowledge of BPMN.

4.2.2 Empirical Comparisons of Security Modeling Notations

In the specific domain of modeling security issues, Stalhane et al. conducted a series of experiments [116, 118, 119] to compare the effectiveness of textual and visual notations in identifying safety hazards during security requirements analysis. Stålhane and Sindre [116] compared misuse cases based on use-case diagrams to those based on textual use cases. The results of the experiment revealed that textual use cases helped to identify more threats related to the computer system and category "wrong patient" than use-case diagrams. This can be explained by the fact that the layout of the textual use case helps the user to focus in the relevant areas which led to better threat identification for these areas. In more recent experiments [117–119] they compared textual misuse cases against UML system sequence diagrams. The experiments revealed that textual misuse cases are better than sequence diagrams when it comes to identifying threats related to functionalities or user behavior. Sequence diagrams outperform textual use cases when it comes to threats related to the system's internal working. The authors concluded that "It is not enough to provide information related to the system's working. It must also be continuously kept in the analyst's focus."

As far as we know, only two studies have investigated the comprehensibility of security risk models. In the first work Hogganvik and Stølen [43] reported two empirical experiments with students to test (a) understanding of the conceptual model of the CORAS

and (b) the use of graphical icons and their effect on the understanding of risk models. The results showed little difference in the correctness of answers using CORAS over UML models, while the participants used less time to complete a questionnaire with the CORAS models than with the UML models. The only difference between the two type of risk models was the presence of graphical CORAS-specific icons. In the second work Grøndahl et al. [35] investigated the effect of textual labels and graphical means (size, color, shape of elements) on the comprehension of risk models. The study involved 57 IT professionals and students and shows that some textual information in graphical models is preferred over purely graphical representation. These works focused on the graphical representation of risk models and leaves open the question of which modeling notation, graphical or textual, is better to represent security risks.

We have started to fill this gap by investigating the actual and perceived effectiveness of textual and visual methods for security risk assessment in two previous empirical studies with MSc students in Security Engineering [55, 58]. Although the two types of methods were similar in terms of actual effectiveness, participants always perceived the visual methods as more effective than the textual methods. For example, Labunets et al. [55] reported that "some of the participants indicated that a visual representation for threat would be better that a tabular one", and in [58] participants emphasized that "the advantage [of graphical method] is the visualization" and that the results obtained with the graphical method would be easy to explain to customer [58, Table III]. In this chapter we explore whether such preference may be explained by the widely held belief that graphical representations are easier to read.

4.3 Study Design

4.3.1 Motivation

In our previous study [57] we conducted a qualitative study with security experts in the ATM domain to investigate the success factors of a security risk assessment. The participants were 20 professionals with 17.5 years of work experience on average and in particular 7 years of experience in risk assessment. As reported in [57, Table 2], among method's success criteria we identified category "Comprehensibility of method outcomes". We have reviewed the experts' statements that were included in this category and discuss them below in order to understand the role of comprehensibility in security risk assessment.

According to some experts "for a method to be successful means that you get the means to reason about your problem and to analyze the information and to extract the results that you want." Indeed, an effective security risk assessment method "must support understanding and communication [of the information]" because the possible shortfall in

the risk assessment process is that "people don't understand each other, so they're using the same words, but they think about totally different things". Besides the common language that should be used throughout risk assessment process, it is also important to have a comprehensive representation: "If you have a good template, it would be easy to understand." Also "you need a definition that lots of people can understand, not just a security expert" in order to have a "basis to share with other stakeholders, and to have the same way of thinking". In fact, you need "to address different stakeholders who look at the risk assessment. And basically you can divide them into two [types]: the ones who need the big picture and the ones who need ... operation knowledge [low level picture] ... The first kind is making the basic decisions and the others for subsequent execution of the results." Some experts believe that "The big picture is effective when you provide usually a graphical representation of it."

4.3.2 Designing Comprehensibility Tasks

The understanding of the results by different stakeholders is one of the main factors for the success of security risk assessment. Different presentations of the same findings might require different levels of cognitive effort to extract the correct information. Hence, we aim to investigate *which risk model representation is more comprehensive for stakeholders from the point of view of extracting correct information about security risks?*

To design a comprehensibility task we reviewed existing works investigating comprehensibility of different notations in requirements engineering [37, 101] and data modeling [17, 90]. In summary, all proposed comprehensibility questions tested the ability of the user to identify (1) an element of a specific type that is in relationship with another element of a different type and (2) an element of a specific type that has multiple relationships with other elements of a different type. We used both approaches to formulate questions for our comprehensibility task as they provide a possibility to investigate the comprehension of different elements of a notation and relations between them.

4.3.3 Task Complexity and Other Factors

We also take into consideration the complexity of the questions, as this may be a significant factor for the risk model comprehensibility. To define this we rely upon the work of Wood [130], according to which a task (or question) complexity is defined by the information cues that need to be processed and the number and complexity of the actions that need to be performed to accomplish the task:

- "Information cues are pieces of information about the attributes of stimulus objects" [130, p. 65];
- "The required acts for the creation of a defined product [output] can be described at any one of several levels of abstraction..." [130, p. 66];

- "Coordinative complexity refers to the nature of relationships between task input and task product. As the number of precedence relationships between acts increases, the knowledge and skill required will also increase..." [130, pp. 68–69].

In the definition of task complexity Wood also used the notion of "product" as a specific entity produced by the task. We do not use this concept because only one product is given to the participants (a risk model) and every question only asks them for one type of element of the risk model. We map other components to the elements of a security risk modeling notation as follows:

- *Information cues (IC)* describe some characteristics that help to identify the desired element of the model. They are identified by a noun. In the sentence "Which are the assets that can be harmed by the unwanted incident *Unauthorized access to HCN*?" the part in italics is an information cue.
- *Required acts (A)* are judgment acts that require selecting a subset of elements meeting some explicit or implicit criteria. For example, in "What is the *highest* consequence?" or "What are the unwanted incidents that *can* occur?" the parts in italics are judgment criteria.
- *Relationships (R)* are relationships between a desired element and other elements of the model that must to be identified in order to find the desired element. They are identified by a verb. In the sentence "the assets that can *be harmed by*", the part in italics is a relationship.

To calculate the *complexity of question i* (QC_i) we extend Wood's formulation as follows:

$$QC_i = |IC_i| + |R_i| + |A_i|, \tag{4.1}$$

where IC_i is the number of information cues presented in question i, R_i is the number of relationships that the participant needs to identify, and A_i is the number of judgments to be performed over a set of elements.

As an example of computing task complexity, consider the following question: "What is the highest possible consequence for the asset "Data confidentiality" that Cyber criminal or Hacker can cause? Please specify the consequence." The question complexity according to formula (4.1) is $3 + (2 + 1) = 6$ because there are three information cues ("Data confidentiality" for the element type "consequence", and "Cyber criminal" and "Hacker" for the element type "threat"), two relationships among them (A "possible consequence for" B and C "can cause" D), and one judgment on the product ("highest possible consequence").

Another possible confounding factor is the complexity of the particular execution of the experiment itself. Therefore, after the comprehension task we asked participants to fill in a post-task questionnaire about their perception of the clarity of the questions and the overall settings and whether the risk model was easy to understand. The aim of the post-task questionnaire is to control for possible effects of the experimental settings on the results as done in previous studies [2,37]. Table 8.1 in Chapter 8 reports the post-task

questionnaire that we proposed to our participants.

4.3.4 Selection of Risk Modeling Notations

There are many different methods for security requirements engineering and risk assessment that use either graphical, or tabular, or mix of two representations. To make the study fair and representative we need to find notations that have similar level of expressiveness and cover the core security concepts used by many international security standards, e.g. ISO/IEC 27000, NIST 800-30, or BSI Standard 100-2 IT-Grundschutz. In this respect, Fabian et al. [22] presented a comprehensive comparison of various security requirements engineering methods based on their conceptual framework that is consistent with the framework by Mayer et al. [68] (see [22, Table 3]). The core concepts that emerged from the studies are *asset, threat, vulnerability, risk, and security control.*

The comparison by Fabian et al. [22] showed that only several methods adopted these concepts, namely tabular *SREP* [71], graphical *CORAS* [61], and model-based *information system security risk management (ISSRM)* approach proposed by Mayer et al. [69]. The ISSMR method initially used i* models to support risk analysis and has been later adapted to by Matulevičius et al. [66] to combine the graphical-based method proposed by Mouratidis and Giorgini [81] *Secure Tropos.*

To the best of our knowledge, the work by Massacci and Paci [65] is the only study that empirically investigated and compared different security methods including Secure Tropos, CORAS, *si∗*, and Security Argumentation. Both CORAS and Secure Tropos methods were empirically evaluated in [65]. The study also included goal-based method *si∗* and problem frame-based method *Security Argumentation.* The results showed that the CORAS is the best method across the four investigated methods.

Further, neither ISSRM nor Secure Tropos provide a comprehensive one-diagram models that provides a global picture of security risk assessment results and that can be compared to a single table summarizing the risk assessment result as provided by NIST's or ISO's standards. In contrast, CORAS has a treatment overview diagram that fits these requirements. Asking the particpants to go over several diagrams would have significantly biased the results against graphical methods.

As tabular representation we used the risk tables provided by the NIST 800-30 [114] standard for security risk assessment. The NIST standard adopts a different table for each step of the security risk assessment process. CORAS similarly comes with a number of different kinds of diagrams. In our study we focused on the NIST table template for adversarial and non-adversarial risk, and the CORAS treatment diagrams, because these two give an overview of the most important elements of the risk assessment. In order to ensure the same expressiveness of the two notations we needed to add three columns to the NIST template to represent impact, asset and security controls, which are usu-

ally documented in different tables. Figure 4.1a shows an example of CORAS treatment diagram related to the risk of a Healthcare Collaborative Network, and Figure 4.1b illustrates the same risks using the NIST table template. The graphical model provides a good visual view of several attacks that can be committed by a "threat". At the same time, tabular model reports all possible attacks (one per line) which requires duplication of the information for the similar attacks with slight difference. However, this redundancy is compensated by simple navigation providing a possibility to look-up the information related to the same notation's concept. The availability of labels with concepts' name may provide a significant benefit comparing to the graphical icons, but Hogganvik and Stølen [43] showed that there is a little difference in the correctness of responses by participants using models with graphical icons from the CORAS notation and UML models that contained textual labels with concepts' names. Moreover, the participants used less time to find response with graphical icons comparing to the UML models with textual labels. Figures 8.1 and 8.2 in Chapter 8 illustrate the full graphical and tabular risk models that we provided to our participants.

4.3.5 Variables

The *independent variable* of our study is the risk model representation which can take one of the values: tabular or graphical. The *dependent variable* is the level of comprehensibility which is measured by assessing the answers of the participants to a series of comprehension questions about the content presented in the risk models. In what follows, we will use the word "task" when referring to the entire exercise of answering all questions. The answers to the questions were evaluated using information retrieval metrics that are widely adopted in the empirical software engineering community for the measurement of model comprehension [2, 37, 100, 101]: *precision*, *recall*, and their harmonic combination, the *F-measure*. Precision represents the correctness of given responses to the question, and recall represents the completeness of the responses. They are calculated as follows:

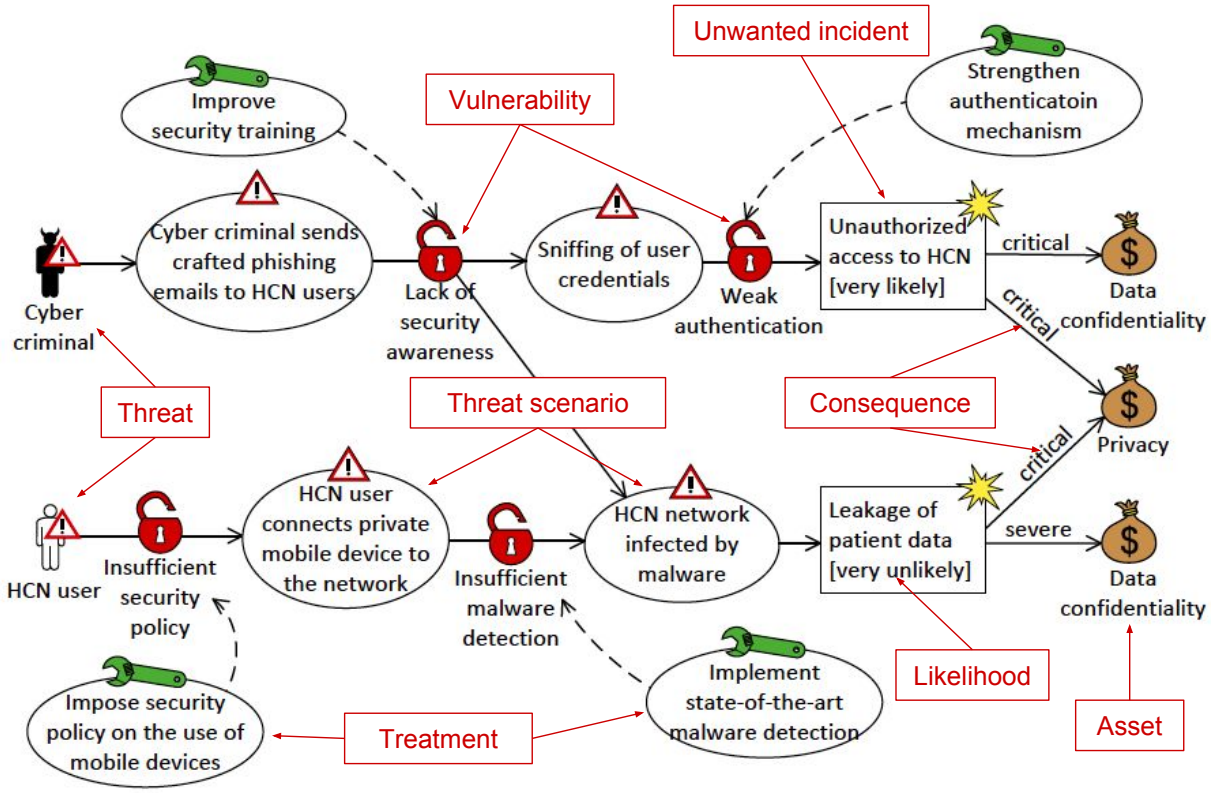
$$precision_{m,s,q} = \frac{|answer_{m,s,q} \cap correct_q|}{|answer_{m,s,q}|}, \tag{4.2}$$

$$recall_{m,s,q} = \frac{|answer_{m,s,q} \cap correct_q|}{|correct_q|}, \tag{4.3}$$

$$F_{m,s,q} = 2 * \frac{precision_{m,s,q} \times recall_{m,s,i}}{precision_{m,s,q} + recall_{m,s,q}}, \tag{4.4}$$

$$F_{m,s} = \text{mean}(\cup_{q \in \{1 \ldots N_{questions}\}} F_{m,s,q}\} \tag{4.5}$$

where $answer_{m,s,q}$ is the set of answers given by participant s to question q when looking at model m, and $correct_q$ is the set of correct responses to question q.



(a) CORAS diagram

Threat Event	Threat Source	Vulnerabilities	Impact	Asset	Overall Likelihood	Level of Impact	Security Controls
Cyber criminal sends crafted phishing emails to HCN users and this leads to sniffing of user credentials.	Cyber criminal	1. Lack of security awareness 2. Weak authentication	Unauthorized access to HCN	Data confidentiality	Very likely	Severe	1. Improve security training. 2. Strengthen authentication mechanism.
Cyber criminal sends crafted phishing emails to HCN users and this leads to sniffing of user credentials.	Cyber criminal	1. Lack of security awareness 2. Weak authentication	Unauthorized access to HCN	Privacy	Very likely	Severe	1. Improve security training. 2. Strengthen authentication mechanism.
Cyber criminal sends crafted phishing emails to HCN users and this leads to that HCN network infected by malware.	Cyber criminal	Lack of security awareness	Leakage of patient data	Privacy	Very unlikely	Critical	Improve security training.
Cyber criminal sends crafted phishing emails to HCN users and this leads to that HCN network infected by malware.	Cyber criminal	Lack of security awareness	Leakage of patient data	Data confidentiality	Very unlikely	Severe	Improve security training.
HCN user connects private mobile device to the network and this leads to that HCN network infected by malware.	HCN user	1. Insufficient security policy 2. Insufficient malware detection	Leakage of patient data	Privacy	Very unlikely	Critical	1. Impose security policy on the use of mobile devices. 2. Implement state-of-the-art malware detection.
HCN user connects private mobile device to the network and this leads to that HCN network infected by malware.	HCN user	1. Insufficient security policy 2. Insufficient malware detection	Leakage of patient data	Data confidentiality	Very unlikely	Severe	1. Impose security policy on the use of mobile devices. 2. Implement state-of-the-art malware detection.

(b) NIST table row entries

Figure 4.1: Fragment of a risk model in graphical and tabular notations

Table 4.1: Experimental Hypotheses

Hyp	Null Hypothesis	Alternative Hypothesis
$H4.1$	No difference between tabular and graphical risk modeling notations in the level of comprehension (as measured by precision, recall, F-measure of answers) when answering comprehension questions.	There is a difference in the level of comprehension between tabular and graphical risk models when answering comprehension questions
$H4.2$	No difference between simple and complex questions in the level of comprehension when answering comprehension questions for both modeling notations	Difference between simple and complex questions in the level of comprehensibility when answering comprehension questions for some modeling notation

Since we want to measure the *level of comprehension* such activity should be performed by keeping the other confounding variable (*time for comprehension*) fixed. Hence we limit the amount of time that can be used to complete the comprehension task. As a consequence, there may be participants which could not answer all questions within the allotted time. We follow the approach in [1] and aggregate all answers to calculate precision and recall for the individual participant.

$$precision_{m,s} = \frac{\sum_{q=1}^{N_{questions}} |answer_{m,s,q} \cap correct_q|}{\sum_{q=1}^{N_{questions}} |answer_{m,s,q}|}, \tag{4.6}$$

$$recall_{m,s} = \frac{\sum_{q=1}^{N_{questions}} |answer_{m,s,q} \cap correct_q|}{\sum_{q=1}^{N_{questions}} |correct_q|}, \tag{4.7}$$

$$F_{m,s} = 2 * \frac{precision_{m,s} \times recall_{m,s}}{precision_{m,s} + recall_{m,s}}. \tag{4.8}$$

A similar function aggregates over participants when reporting $precision_{m,q}$ and $recall_{m,q}$ for each question q.

4.3.6 Hypotheses

The main objective of our study was to compare the effectiveness of tabular and graphical approached for risk modeling in extracting information about security risks from the models ($RQ4.1$). Additionally, we wanted to investigate if the complexity of comprehension task affects participation' comprehension of risk models. We formulated the alternative two-way hypotheses as there is no consensus about the superiority of one type of notation over the other in the literature (see Section 4.2), and therefore, we did not make any assumptions in this regard. For example, [118] and [43] report opposite results on the superiority of the textual and graphical notation for the comprehension of use cases. Thus, the null and alternative hypotheses were formulated as presented in Table 4.1.

4.3.7 Experimental Design

In the first study (study 4.1) we chose a *between-subjects design* with one factor (risk modeling notation) and two treatments (graphical and tabular risk models) to avoid interference between the treatments [62, Ch. 5]. The participants were randomly assigned to one of the two treatments and worked individually. Each experiment that we executed followed the same design. The graphical and tabular risk models provided to the participants are presented in Chapter 8 in Figures 8.1 and 8.2 respectively. The material used during the experiment is available online (e.g., risk models and tutorial slides).[1]

The experiments consist of three main phases:

- *Training phase.* All participants attend a short 10 min presentation about both types of risk models and the application scenario. Then they answer a short demographics and background questionnaire.
- *Application phase.* During this phase the participants are asked to review proposed graphical or tabular risk models of the application scenario and complete the task which contains 12 comprehension questions. The order of the questions in the task was randomized for each participant. Moreover, the participants are randomly assigned to Group 1 or Group 2 so that half of them answer questions related to the graphical risk model, and the other half respond to questions on the tabular risk model. We ask participants to complete the task in 40 minutes. All necessary materials, like risk model diagrams or tables and tutorial slides, are provided to the participants in electronic form at the beginning of the task. After completion of the task, the participants answer a post-task questionnaire.
- *Evaluation phase.* Researchers independently check the responses of the participants and code correct and wrong answers to each comprehension question based on the predefined list of correct responses.

Inspired by similar studies [17, 37, 44], for the second study (study 4.2) we chose a *within-subjects design* with two factors (risk modeling notation and application scenario) and two levels for each factor. This allowed us to collect participants' level of comprehension of both risk models. To mitigate a possible effect of the treatments' order on the experimental results we used a Latin square. Table 4.2 summarizes the experimental design that we adopted. The participants were randomly assigned to one of the four groups and worked individually. The graphical and tabular risk models provided to the participants were similar to the ones used in the first study with several small changes. We have made available online the risk models and tutorial slides that we used in the second study.[2]

[1] https://securitylab.disi.unitn.it/doku.php?id=validation_of_risk_and_security_requirements_methodologies

[2] https://securitylab.disi.unitn.it/doku.php?id=unitn-comprehensibility-exp-2015

Table 4.2: Experimental design of the second study

Each group applied one of the method on a scenario and then the second method on the remaining scenario (OB=Online Banking scenario; HCN=Health Care Network scenario; Tab=Tabular risk modeling notation; Graph=Graphical risk).

Session	Group 1	Group 2	Group 3	Group 4
Session 1	Tab; OB	Tab; HCN	Graph; OB	Graph; HCN
Session 2	Graph; HCN	Graph; OB	Tab; HCN	Tab; OB

Table 4.3: Comprehension questionnaire design

Half of the answers require no judgment and combine 1 or 2 information cues connected by 1 or 2 relationships. The other half of the questions have the same combination of information cues and relationships augmented by the judgment element. There are no question with one information cue and two relationships as this combination is impossible.

	One Relationship	Two Relationships
One information Cue	2 questions	-
One Information Cue + Judgment	2 questions	-
Two information Cues	2 questions	2 questions
Two Information Cues + Judgment	2 questions	2 questions

The experimental procedure of the second study is similar to the one reported previously, with one difference. Basically, each session of the second study is the application phase. Therefore, in the second study we have two consecutive application phases (Session 1 and Session 2) of about 40 minutes each. To mitigate the learning effect in Session 2 each participant receives a treatment different from the one that he received in Session 1. Section 4.5.4 will provide statistical verification that there were no significant differences between the results of the two sessions and between the results of the two application scenarios.

Comprehension Questionnaire Revision. The results of the first study revealed a statistically significant effect of task complexity on the participants' comprehension of the risk models. Thus, we revised the comprehensibility questions for our second study with the focus on the task complexity to better investigate *RQ*4.2. Table 4.3 presents the distribution of the questions by the number of information cues, relationships and judgments present in the question. Table 8.3 in Chapter 8 reports the comprehension questionnaire for the graphical risk model in the second study. Similar to the first study these questions were reviewed by independent researchers from SINTEF who are the experts in the graphical risk modeling notation. The questions for the textual risk model are the same but the names used to denote the elements and relations are instantiated to the textual risk modeling notation.

4.3.8 Selection of Application Scenarios

In the first study we used an application scenario developed by IBM about the Healthcare Collaborative Network (HCN). HCN is a health information infrastructure for interconnecting and coordinating the delivery of information to participants in the collaborative network electronically.

In the second study in order to avoid learning effects between two application sessions we used two different application scenarios. In addition to the HCN scenario, we used an Online Banking scenario developed by Poste Italiane, describing online banking services provided by Poste Italiane's division through a home banking portal, a mobile application and prepaid cards.

The graphical risk models for the two application scenarios were developed by independent researchers from the Norwegian research institute SINTEF who are the designers of the CORAS graphical risk modeling notation in the framework of the EMFASE project. We developed the corresponding tabular risk models. After the models were developed, together with experts from SINTEF we checked that the models are conceptual copies of one another to the extent that the two different notations allow this.

For each risk model we developed the comprehension questionnaire. The questionnaires were reviewed by the researchers from SINTEF. In cooperation with the designers from SINTEF we developed the list of correct responses. Tables 8.2 and 8.3 in Chapter 8 report the comprehension questionnaire for the graphical risk model for both studies. The questions for the textual risk model are identical but for the names used to denote the elements and relations that are instantiated to the textual risk modeling notation.

4.3.9 Analysis Procedure

We test the null hypothesis $H4.1_0$ using an unpaired statistical test in the first study as we have a between-subjects design, and a paired statistical test in the second study because of a within-subjects design. Distribution normality is checked by the Shapiro–Wilk test. If our data are normally distributed we use an unpaired t-test to compare comprehension of independent groups in the first study and paired t-test to compare the comprehensibility of matched groups in the second study; otherwise we use their non-parametric analogs, the MW and Wilcoxon tests respectively.

We investigate the effect of task complexity and test the null hypothesis $H4.2_0$ using the Wilcoxon test for non-normal distribution. We have paired data because we investigate the difference in responses to questions with different complexity level obtained from the same participant.

We also use interaction plots to check the possible effects of co-factors on the dependent variable. If the plot reveals any interaction between co-factors and the treatment we also use a permutation test for two-way ANOVA to check whether this interaction is

Table 4.4: Participants Distribution to Treatments – Study 4.1

In total 36 participants completed the comprehension task using the graphical risk model and 33 participants used the tabular notation.

Experiment	Graph	Tabular	Total
4.1. UNITN-MSC	18	17	35
4.1. PUCRS-MSC	6	7	13
4.1. PUCRS-BSC	12	9	21
Total	36	33	69

statistically significant. The post-task questionnaire is used to control for the effect of the experimental settings and the documentation materials.

We adopt 5% as a threshold of α (i.e., the probability of committing Type-I error). To report the effect size of observed differences between treatments we used Cohen's d with the following thresholds: *negligible* for $|d| < 0.2$, *small* for $0.2 \leq |d| < 0.5$, *medium* for $0.5 \leq |d| < 0.8$, and *large* for $|d| \geq 0.8$. To run statistical tests we used `RStudio`[3].

4.4 Study Realization

4.4.1 Experiments Execution

Table 4.4 summarizes the experimental set-up for the first study. The first experiment was conducted at the University of Trento in the fall semester of 2014 as part of the Security Engineering course. The participants were 35 MSc students in Computer Science. The experiment took place in a single computer laboratory. The experiment was presented as a laboratory activity and only the high-level goal of the experiment was mentioned; the experimental hypotheses were not provided so as not to influence the participants but they were informed about the experimental procedure. At the end of the experiment we had a short discussion on the experiment's procedure and on the two modeling notations.

The same settings were maintained in two replicated experiments which were executed at the PUCRS University in Porto Alegre, Brazil. The first replication involved 13 MSc students enrolled in the Computer Science program. The second one involved 27 BSc students attending the Information Systems course taught at the Computer Science department. Both replications took place in a single computer laboratory.

Six participants failed to complete the task and we discarded their results: one participant answered the question in Portuguese instead of English and they were not related to the model, other participants did not provide responses based on the model.

Table 4.5 summarizes the experimental set-up for the second study. The first experiment was conducted in Cosenza at Poste Italiane cyber-security lab (a large corporation)

[3] `www.rstudio.com`

Table 4.5: Participants Distribution to Treatments – Study 4.2

In total we had 83 participants who were randomly assigned to one of four groups. The description of the groups see in Table 4.2. each group answered questions on a scenario described in one risk modeling notation and then questions on a different scenario on the other risk modeling notation.

Session	Group 1	Group 2	Group 3	Group 4	Total
4.2. POSTE	12	9	10	10	41
4.2. UNITN	12	10	10	10	42

in September 2015. The participants were 52 MSc/MEng graduates attending a professional master course in Cybersecurity. The experiment took place in a single computer laboratory. The experiment was presented as an entry evaluation activity for the course and only the high-level goal of the experiment was revealed. The participants were instructed about the experimental procedure.

The same settings were kept in the replication conducted at the University of Trento in October 2015 as part of the Security Engineering course. The replication involved 51 MSc students in Computer Science. The experiment was presented as a laboratory activity.

There were some participants who failed to complete both sessions, i.e. they finished the task at home, or had a problem with the SurveyGizmo platform and restarted their task[4]. We removed the responses of these participants from our dataset to eliminate the bias created by the varying time. In total we discarded 11 participants from the first experiment (21%) and 9 participants from the second one (18%) which allowed us to keep a significant number of participants without compromising the internal validity of the experiment.

4.4.2 Demographics

Table 4.6 summarizes the demographic information about the participants of our experiments for the first study. Most participants (75%) reported that they had working experience. With respect to security knowledge most participants had limited expertise. In contrast, they reported good general knowledge of modeling languages: software engineering courses taught at both universities are compulsory and included several lectures on UML and other graphical modeling notations. The participants only had very basic knowledge of the application scenario.

Table 4.7 summarizes the demographic information about the participants of our experiments for the second study. Most participants (52%) reported that they had working experience. The participants of the second study had slightly better security knowledge

[4]When a participant by mistake closes the web page with the task in SurveyGizmo she loses the session and cannot restore it and must restart from scratch. From the platform perspective she has used the same amount of time of other participants, but in practice might have had significantly more time.

Table 4.6: Demographic Statistics – Study 4.1

The participants were 35 Italian MSc students attending a Security Engineering course at the University of Trento, 13 MSc and 21 BSc students studying Computer Science the PUCRS University in Porto Alegre, Brazil.

Variable	Scale	Mean/Med.	Distribution
Age	Years	25.8 (mean)	45% were 19–23 yrs old; 38% were 24–29 yrs old; 19% were 30–46 yrs old
Gender	Sex		78% male; 22% female
Work experience	Years	3.9 (mean)	25% had no experience; 43% had 1–3 yrs; 15% had 4–7 yrs; 17% had >7 yrs
Expertise in security	0(Novice)–4(Expert)	1 (median)	29% novices; 49% beginners; 17.5% competent users; 4.5% proficient
Expertise in modeling languages	0–4	2 (median)	11.5% novices; 21.5% beginners; 54% competent users; 10% proficient users; 3% experts
Expertise in HCN	0–4	0 (median)	67% novices; 23% beginners; 10% competent users

Table 4.7: Demographic Statistics – Study 4.2

The participants were 42 Italian MSc/MEng graduates attending a professional master in cybersecurity in Cosenza organized by Poste Italiane, a large corporation, and 41 MSc students attending a security engineering course at the University of Trento.

Variable	Scale	Mean/Med.	Distribution
Age	Years	26.4 (mean)	25% were 21–23 yrs old; 55% were 24–29 yrs old; 20% were 30–40 yrs old
Gender	Sex		75% male; 25% female
English level	A1–C2		1% Elementary (A1); 5% Pre-Intermediate (A2); 37% Intermediate (B1); 31% Upper-Intermediate (B2); 14% Advanced (C1); 11% Proficient (C2)
Work experience	Years	1.3 (mean)	49% had no experience; 39% had 1–3 yrs; 11% had 4–7 yrs; 1% had >7 yrs
Expertise in security	0(Novice)–4(Expert)	1 (median)	19% novices; 52% beginners; 19% competent users; 5% proficient; 5% experts
Expertise in modeling languages	0–4	2 (median)	15% novices; 33% beginners; 36.5% competent users; 13% proficient users; 2.5% experts
Expertise in online banking	0–4	0 (median)	73% novices; 21% beginners; 4% competent users; 1% proficient users; 1% experts
Expertise in HCN	0–4	0 (median)	80% novices; 19% beginners; 1% experts

and slightly worse knowledge of modeling languages compared to the participants of the first study (see Table 4.6). They also had very basic knowledge of the application domains.

4.5 Experimental Results

In this section we report the results obtained in two studies and its analysis. The results of preliminary analysis with Shapiro–Wilk test showed that our dependent variable (precision and recall) was not normally distributed. Thus, in $RQ4.1$ we proceeded with a non-parametric MW test for the results of the first study as it has a between-subjects design and with Wilcoxon test for the second study because it has a within-subjects design. In $RQ4.2$ we used Wilcoxon test as we compare the responses to questions with different complexity but from the same participant, and therefore, our data were paired.

4.5.1 RQ4.1: Effect of Risk modeling notation on Comprehension

Tables 4.8 and 4.9 report descriptive statistics for precision and recall based on the results of application phase across experiments of the first and second study respectively. As can be seen, in the first study the answers to the questions on the tabular risk model demonstrated 7% better average precision and 22% better average recall over the questions posed on the graphical risk model. In the second study we got similar results: the responses to the questions on the tabular risk model showed an overall 13% better precision and an overall 30% better recall over the responses given with the graphical risk model. We also report precision and recall by questions in Tables 8.4 and 8.5 in Appendix.

Figure 4.2 presents precision and recall of participants' responses to the comprehension task in the two studies. Participants who used tabular risk model showed better precision and recall of responses than the participants who used a graphical model. Tables 4.8 and 4.9 support this observation. When looking at individual experiments we can observe that in the first study the participants of experiment PUCRS-BSC demonstrated the least difference in precision. A possible reason can be language issue as the participants were BSc students from Brazil speaking Portuguese and may have problems with understanding English text.

The $H4.1_0$ is tested with Wilcoxon and MW tests and the results presented in Table 4.10. The tests revealed a statistically significant difference in precision and recall for most of the experiments with effect size ranging from small to very large except PUCRS-BSC where we obtained p-value > 0.05. Only for overall recall of the first study Levene's test returned p-value <0.05 which means that sample does not meet homogeneity of variance assumption required by MW test. To validate its result we run Kruskal-Wallis test that can be used instead of MW test and does not require homogeneity of variance. The test returned p-value $= 1.2 * 10^{-5}$ and confirmed the findings of MW test. Overall, we can conclude that the tabular risk modeling notation is more effective in supporting comprehension of security risks than the graphical one.

Table 4.8: Descriptive Statistics of Precision and Recall by Modeling Notation – Study 4.1

For both precision over all questions and recall over all questions the tabular risk model was easier to comprehend than the graphical one within each experiment and overall across the three experiments.

	Tabular			Graphical		
	Mean	Median	sd	Mean	Median	sd
Precision						
4.1. UNITN-MCS	0.90	0.92	0.06	0.84	0.88	0.11
4.1. PUCRS-MCS	0.82	0.87	0.12	0.70	0.74	0.10
4.1. PUCRS-BSC	0.81	0.90	0.15	0.80	0.83	0.13
Overall	0.86	0.92	0.11	0.80	0.84	0.12
Recall						
4.1. UNITN-MCS	0.89	0.89	0.07	0.75	0.78	0.15
4.1. PUCRS-MCS	0.89	0.93	0.09	0.61	0.66	0.11
4.1. PUCRS-BSC	0.89	0.96	0.12	0.75	0.79	0.17
Overall	0.89	0.89	0.09	0.73	0.76	0.16

Table 4.9: Descriptive Statistics of Precision and Recall by Modeling Notation – Study 4.2

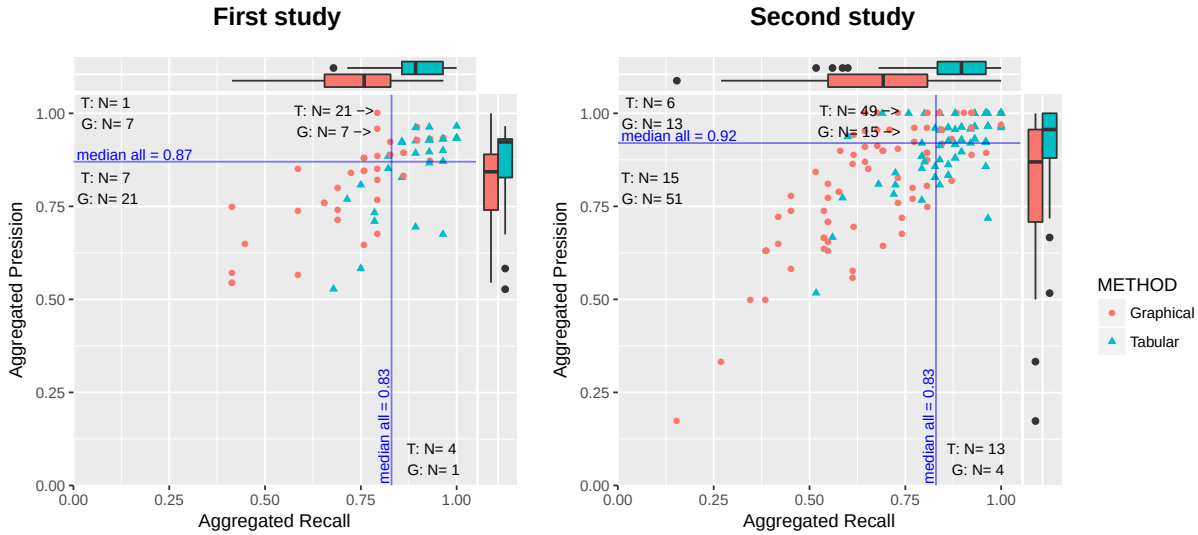
For both precision and recall over all questions the tabular risk model was easier to comprehend than the graphical one within each experiment and overall across the two experiments.

	Tabular			Graphical		
	Mean	Median	sd	Mean	Median	sd
Precision						
4.2. POSTE	0.92	0.96	0.09	0.80	0.88	0.19
4.2. UNITN	0.93	0.95	0.09	0.84	0.86	0.14
Overall	0.92	0.96	0.09	0.82	0.87	0.17
Recall						
4.2. POSTE	0.87	0.88	0.11	0.64	0.65	0.19
4.2. UNITN	0.89	0.91	0.11	0.71	0.72	0.17
Overall	0.88	0.90	0.11	0.68	0.69	0.18

4.5.2 RQ4.2: Effect of Task Complexity on Comprehension

Figures 4.3a and 4.3b compare the distribution of precision and recall of the participants' responses to full comprehension task (Q1–Q12) (left) and only to the complex questions (right), namely question complexity level > 2.

There is a significant difference in recall of the responses to the complex questions between tabular and graphical risk models. In the first study 76% of the participants who used the tabular risk model achieved recall better than or equal to the overall median value, whilst only 28% of the participants who used the graphical risk model passed the recall threshold. In the second study we observed bigger difference: 80% and 23% of the participants passed therecall threshold in tabular and graphical group respectively.



For both studies participants using a tabular risk model showed a much better significant recall than the graphical one (see the number of points to the left of median bar and the non overlapping boxplots on the top of the diagrams). The participants using a graphical model have a slightly lower significant precision than participants using tabular models as can be seen from the number of points below the median bar and the boxplots on the right of the diagrams.

Figure 4.2: Distribution of Participants' Precision and Recall by Modeling Notation

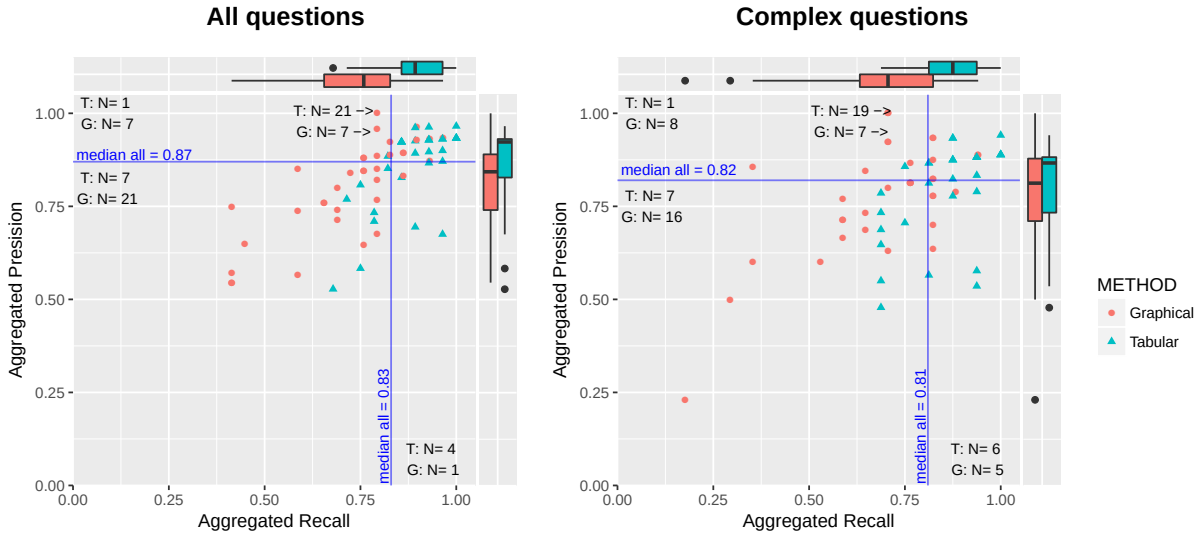
Table 4.10: RQ4.1 – Summary of Experimental Results by Modeling Notation

The results of Wilcoxon test for the first study and MW test for the second study revealed showed that tabular risk modeling notation are statistically easier to comprehend as measured by both in precision (small-medium effect) and in recall (large-very large effect) at the 5% confidence level.

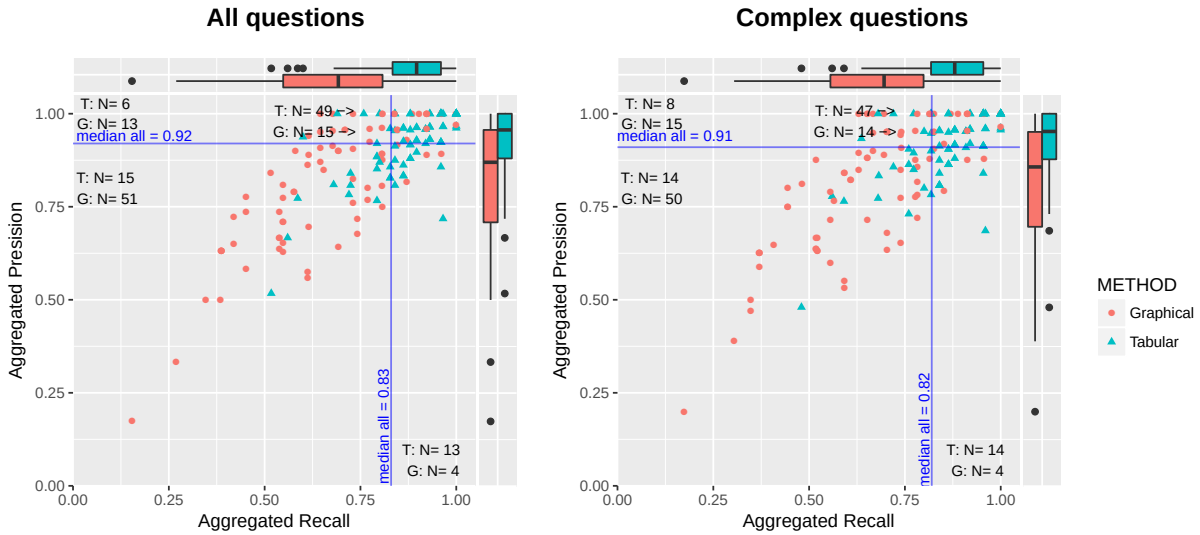
	Experiment	#part.	#obs.	μ_T-μ_G	σ	p-value	Cohen's	d
Precision	4.1. UNITN-MCS	35	35	0.06	0.12	**0.024**	Small	0.49
	4.1. PUCRS-MCS	13	13	0.13	0.18	**0.046**	Medium	0.71
	4.1. PUCRS-BSC	21	21	0.01	0.22	0.66	Negligible	0.06
	4.2. POSTE	41	82	0.12	0.15	$\mathbf{6.7 \cdot 10^{-5}}$	Medium	0.79
	4.2. UNITN	42	84	0.09	0.12	$\mathbf{4.1 \cdot 10^{-6}}$	Large	0.81
	Study 4.1: Overall	69	69	0.06	0.17	**0.018**	Small	0.32
	Study 4.2: Overall	83	166	0.11	0.13	$\mathbf{1.9 \cdot 10^{-8}}$	Medium	0.79
Recall	4.1. UNITN-MCS	35	35	0.14	0.14	**0.002**	Large	0.95
	4.1. PUCRS-MCS	13	13	0.27	0.15	**0.001**	Very large	1.87
	4.1. PUCRS-BSC	21	21	0.15	0.21	**0.054**	Medium	0.7
	4.2. POSTE	41	82	0.23	0.16	$\mathbf{1.9 \cdot 10^{-9}}$	Very large	1.46
	4.2. UNITN	42	84	0.18	0.14	$\mathbf{5.7 \cdot 10^{-9}}$	Very large	1.25
	Study 4.1: Overall	69	69	0.16	0.17	$\mathbf{5.0 \cdot 10^{-6}}$	Large	0.98
	Study 4.2: Overall	83	166	0.2	0.15	$\mathbf{4.1 \cdot 10^{-13}}$	Very large	1.35

In the case of precision the gap in comprehension is reduced: in the first study 67% and 39% of the participants who used respectively tabular and graphical risk models passed the threshold. In the second study the difference is smaller and these proportions were 66% and 34% for tabular and graphical risk models respectively.

To better investigate this effect, we used the interaction plots between precision, recall, and questions' complexity. Figures 4.4a and 4.4b shows that there is no significant interaction between precision, recall and risk modeling notation.



(a) Study 4.1

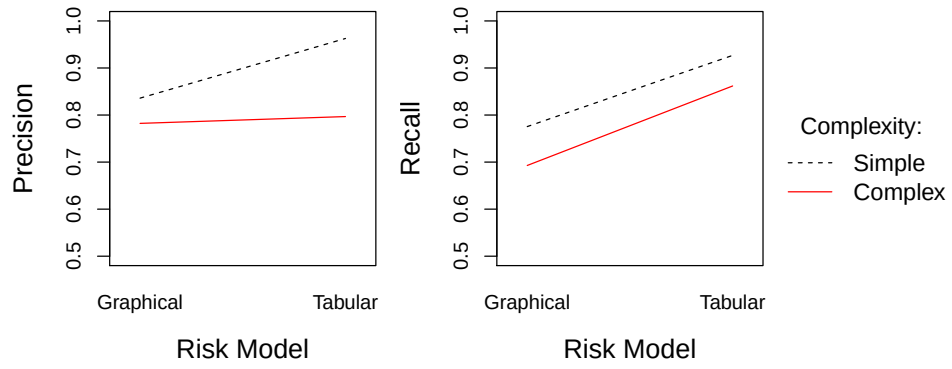


(b) Study 4.2

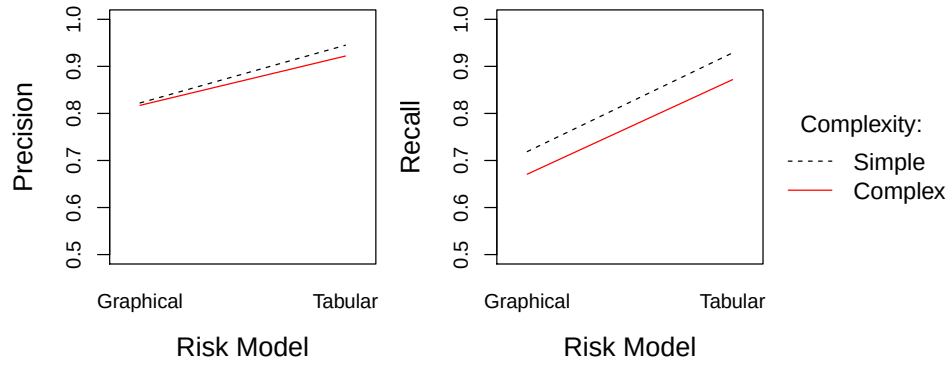
For both simple and complex questions participants using a tabular risk model have better recall than the graphical one. There is a significant difference in precision across simple and complex questions. The participants using a graphical model have a lower precision than participants using tabular models as can be seen from the larger number of points below the median bar and the boxplot on the right of the diagrams.

Figure 4.3: Distribution of Participants' Precision and Recall by Task Complexity

For both simple and complex questions the tabular risk model has better recall than the graphical one and this holds for both studies. The difference in precision is significant only in the first study, where tabular risk model showed significantly better precision for simple questions (0.96 as mean value) over the complex ones (0.80). In the second study for both risk modeling notations there is no significant difference in precision between simple and



(a) Study 4.1



(b) Study 4.2

There is no significant interaction between precision, recall, and risk modeling notation. Only for simple questions participants using the tabular notation performed significantly better albeit for a small effect.

Figure 4.4: Interaction Among Risk Modeling Notation and Task Complexity

complex questions. As there is no major interaction between risk model notation and either precision or recall, we can simply use the F-measure as an aggregated measure of participants' comprehension for further co-factor analysis and for answering $RQ4.2$.

To make this analysis more precise we calculate the F-measure by aggregating it by questions' complexity, so that $F_{m,s,\ell}$ is the mean value for participant s using risk model m over all questions q with complexity level ℓ. We aggregate the levels as $\ell = 2$ and $\ell > 2$ (see complexity levels in Tables 8.4 and 8.5 in Appendix). The formulation is essentially identical to (4.5) except that q only ranges over the questions with complexity ℓ.

Tables 4.11 and 4.12 presents the descriptive statistics for F-measure of simple and complex questions for tabular and graphical models in two studies. In both studies participants' obtained better F-measure for simple questions in comparison to the complex ones. Interesting fact that the participants of the experiment PUCRS-MCS in the first

Table 4.11: Descriptive Statistics of F-measure by Task Complexity – Study 4.1

In the first study F-measure of simple questions was significantly higher than of complex questions and this is true for both risk modeling notations. Only in experiment PUCRS-MSC when participants used graphical risk model the difference in F-measure between simple and complex questions was smaller (0.03) than in the other experiments.

		Simple			Complex		
		Mean	Median	sd	Mean	Median	sd
Tabular	4.1. UNITN-MCS	0.98	1.00	0.04	0.83	0.88	0.10
	4.1. PUCRS-MCS	0.90	1.00	0.17	0.82	0.86	0.13
	4.1. PUCRS-BSC	0.91	1.00	0.17	0.81	0.84	0.13
	Overall	0.94	1.00	0.12	0.82	0.86	0.11
Graphical	4.1. UNITN-MCS	0.85	0.86	0.15	0.75	0.80	0.17
	4.1. PUCRS-MCS	0.67	0.66	0.18	0.64	0.65	0.13
	4.1. PUCRS-BSC	0.81	0.85	0.23	0.74	0.79	0.14
	Overall	0.80	0.83	0.19	0.73	0.79	0.15

Table 4.12: Descriptive Statistics of F-measure by Task Complexity – Study 4.2

In the second study still there was a difference in F-measure in favor of simple questions over the complex ones, but it was smaller for tabular risk model and same for the graphical one. In experiment UNITN the participants who used graphical risk model obtained same mean F-measure for simple and complex questions (0.76).

		Simple			Complex		
		Mean	Median	sd	Mean	Median	sd
Tabular	4.2. POSTE	0.93	1.00	0.20	0.89	0.90	0.09
	4.2. UNITN	0.94	1.00	0.15	0.90	0.91	0.09
	Overall	0.93	1.00	0.17	0.89	0.90	0.09
Graph.	4.2. POSTE	0.76	0.86	0.27	0.70	0.75	0.18
	4.2. UNITN	0.76	0.86	0.26	0.76	0.79	0.15
	Overall	0.76	0.86	0.26	0.73	0.77	0.17

study obtained small difference (0.03) and UNITN in the second study showed no difference in F-measure of simple and complex questions when respond using graphical risk model.

The $H4.2_0$ is tested with Wilcoxon test and the results reported in Table 4.13. Overall the results revealed small but statistically significant difference in favor of simple questions. The difference is significant in most of the experiments when participants' used tabular risk model but not for graphical one. We can conclude that tabular notation is more prone to the effect of task complexity comparing to the graphical notation.

In Section 8.1 we report the additional information showing the effect of different task complexity elements (IC, R, and J) on F-measure by mean of interaction plots.

4.5.3 Post-task Questionnaire

To control the effect of the experiment settings on the results, we analyzed participants' feedback collected with post-task questionnaire after the application task. Tables 4.14a and 4.14b present descriptive statistics of the responses to post-task questionnaire of the first and second studies respectively. Responses are on a five-category Likert scale from 1 (strongly disagree) to 5 (strongly agree). Overall, for both tabular and graphical risk

Table 4.13: RQ4.2 – Summary of Experimental Results by Tasks' Complexity

The results of Wilcoxon test for tabular risk model revealed a statistically significant difference in F-measure in favor of simple questions ($\mu_C \leq \mu_S$). Only for PUCRS-MSC and PUCRS-BSC experiments the test returned p-value > 0.05. The results for graphical risk modeling notation is less convincing as only the experiment UNITN in the first study and overall for the first study we obtained significant results and only for a small effect.

	Experiment	#part.	#obs.	μ_C-μ_S	σ	p-value	Cohen's	d
Tabular	4.1. UNITN-MCS	17	17	-0.14	0.08	$\mathbf{1.5 \cdot 10^{-5}}$	Very large	1.69
	4.1. PUCRS-MCS	7	7	-0.08	0.23	0.30	Small	0.36
	4.1. PUCRS-BSC	9	9	-0.10	0.23	0.055	Small	0.45
	4.2. POSTE	41	41	-0.04	0.24	**0.0003**	Negligible	0.19
	4.2. UNITN	42	42	-0.04	0.20	**0.002**	Negligible	0.18
	Study 4.1: Overall	33	33	-0.12	0.17	$\mathbf{1.1 \cdot 10^{-5}}$	Medium	0.68
	Study 4.2: Overall	83	83	-0.04	0.22	$\mathbf{6.4 \cdot 10^{-6}}$	Negligible	0.18
Graphical	4.1. UNITN-MCS	18	18	-0.09	0.23	**0.03**	Small	0.41
	4.1. PUCRS-MCS	6	6	-0.03	0.25	1.00	Negligible	0.11
	4.1. PUCRS-BSC	12	12	-0.07	0.30	0.15	Small	0.23
	4.2. POSTE	41	41	-0.06	0.36	0.08	Negligible	0.16
	4.2. UNITN	42	42	0.00	0.33	0.41	Negligible	-0.00
	Study 4.1: Overall	36	36	-0.07	0.27	**0.01**	Small	0.28
	Study 4.2: Overall	83	83	-0.03	0.35	0.06	Negligible	0.08

Table 4.14: Post-task Questionnaire Results

For both modeling notations participants agreed that settings were clear, tasks were reasonable, and documentation was clear and sufficient. Scale from 1 (strongly disagree) to 5 (strongly agree).

(a) Study 4.1

	Tabular			Graphical		
Q#	Mean	Median	sd	Mean	Median	sd
Q1	4.67	5.00	0.54	4.67	5.00	0.54
Q2	3.88	4.00	1.05	3.88	4.00	1.05
Q3	4.18	4.00	0.68	4.18	4.00	0.68
Q4	4.00	4.00	0.75	4.00	4.00	0.75
Q5	4.00	4.00	0.83	4.00	4.00	0.83
Q6	4.27	4.00	0.76	4.27	4.00	0.76
Q7	4.33	4.00	0.82	4.33	4.00	0.82
Q8	4.30	4.00	0.77	4.30	4.00	0.77
Q9	Yes (64%) / No (36%)			Yes (50%) / No (50%)		

(b) Study 4.2

	Tabular			Graphical		
Q#	Mean	Median	sd	Mean	Median	sd
Q1	4.22	4.00	0.83	4.22	4.00	0.83
Q2	3.86	4.00	0.84	3.86	4.00	0.84
Q3	4.10	4.00	0.77	4.10	4.00	0.77
Q4	3.93	4.00	0.82	3.93	4.00	0.82
Q5	3.92	4.00	0.80	3.92	4.00	0.80
Q6	3.98	4.00	0.78	3.98	4.00	0.78
Q7	4.02	4.00	0.87	4.02	4.00	0.87
Q8	4.04	4.00	0.77	4.04	4.00	0.77
Q9	Yes (45%) / No (55%)			Yes (39%) / No (61%)		

models participants concluded that the time allocated to complete the task was enough ($Q1$). Participants who used the tabular risk model were more confident in the adequacy of allocated time than participants who used the graphical risk model. They found the objectives of the study ($Q2$) and the task ($Q3$) clear. In general, the participants were confident that the comprehension questions were clear ($Q4$) and they did not experience difficulty in answering the comprehension questions ($Q5$). Also, neither group experienced significant difficulties in understanding ($Q6$) and using electronic versions ($Q7$) of risk model tables or diagrams. The online survey tool was also easy to use ($Q8$).

Since we provided participants with electronic versions of the tabular and graphical risk models, we decided to investigate whether the participants used search/filtering information in tables and diagrams. In the first study most of the participants (64%) who used tabular risk models also used search or filtering information in a browser or MS

Excel, while only half of the participants who used the graphical risk model used search in PDF format. In the second study this ratio was 21% less for participants who used the tabular risk model and 11% lower for participants who used the graphical risk model.

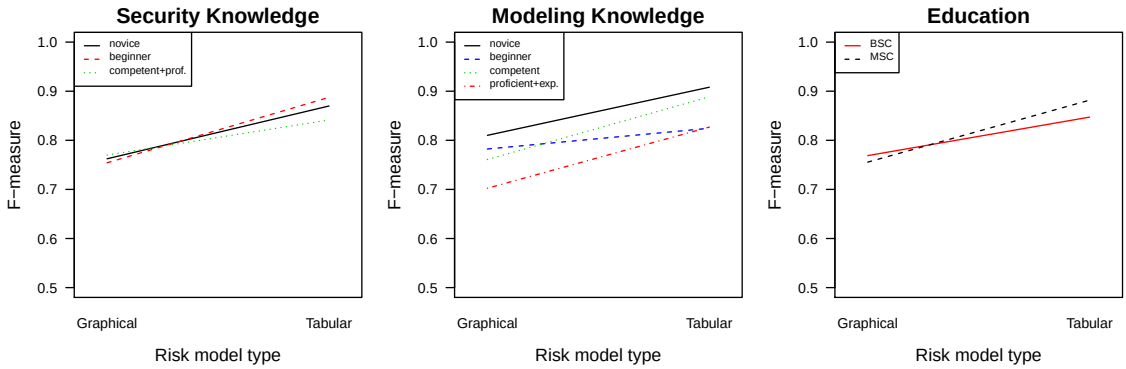
4.5.4 Co-factor Analysis

We investigated the effect of co-factors on the dependent variable through interaction plots. We considered co-factors like education degree (BSc or MSc), working experience, experience in security and privacy projects or initiatives, and level of expertise in security, modeling languages, and in the domain. In the first study only an handful number of participants reported their knowledge as "proficient user" in Security, and therefore we merged this category with the category "competent user". For the same reason we merged the category "expert" in Modeling with the category "proficient user". Similarly, in the second study we had small number of participants who reported their knowledge as "expert" in either Security or Modeling we merged this category with the category "proficient user".

Figure 4.5a shows the interaction plots between the F-measure by modeling notation (graphical vs. tabular) and education degree, security knowledge, or modeling knowledge for the first study. The results of permutation test for two-way ANOVA showed that these interactions are not statistically significant. The test returned $p = 0.55$ for security knowledge vs risk modeling notation, $p = 0.74$ for modeling knowledge vs risk modeling notation, and $p = 0.42$ for education degree vs risk modeling notation. Thus, we did not observe a statistically significant interaction between factors and dependent variable.

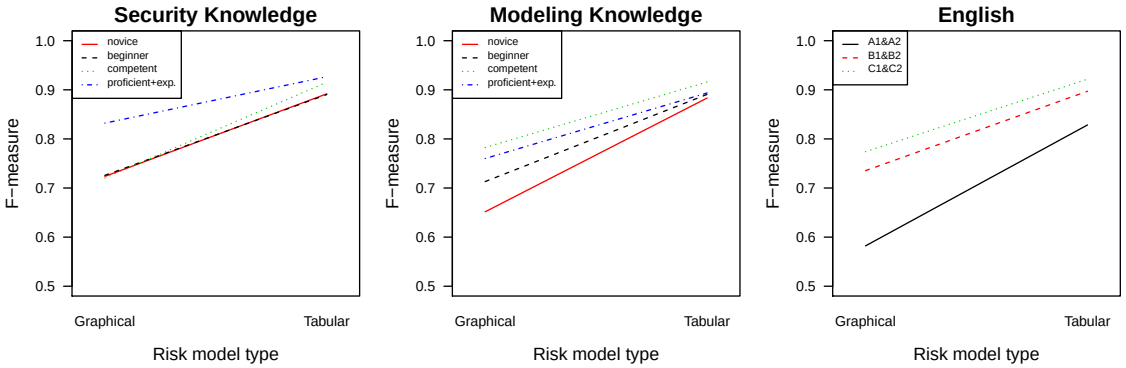
In the experiments of the second study we considered co-factors like knowledge of English, working experience, experience in security and privacy projects or initiatives, level of expertise in security, modeling languages and in the domain. Figure 4.5b shows the interaction plots between the F-measure by modeling notation (graphical vs. tabular) and level of English, security knowledge, or modeling knowledge. The results of permutation test for two-way ANOVA showed that these interactions are not statistically significant. The test returned $p = 0.95$ for the security knowledge level and risk modeling notation, $p = 0.56$ for the modeling knowledge level and risk modeling notation, and $p = 0.38$ for the level of English and risk modeling notation. Thus, in the second study we did not observe a statistically significant effect of co-factors on the experimental results.

Learning Effect in Study 4.2: We investigated a possible learning effect that may be caused by between-subjects design. Figure 4.6 shows the interaction plots between F-measure by modeling notation and scenario and session. The results of permutation test for two-way ANOVA test show that there are no statistically significant interactions. The test returned $p = 0.88$ for the scenario and risk modeling notation and $p = 0.96$ for the session and risk model type.



(a) Study 4.1

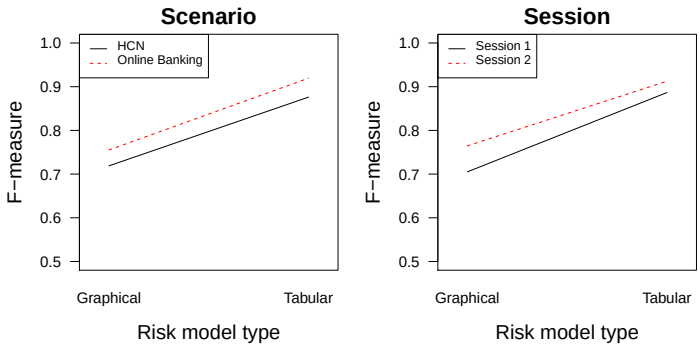
Better expertise corresponds to obviously better results but otherwise security and modeling expertise do not interact with the modeling notation. There is only a limited interaction for participants who are just competent in modeling notation but this is not confirmed by either novices or experts. A permutation test for two-way ANOVA did not reveal any statistically significant interaction.



(b) Study 4.2

Once again better expertise corresponds to better results but otherwise security and modeling expertise do not have major interactions with the modeling notation. The difference in performance due to expertise is smaller for participants using the tabular notation. The permutation test for two-way ANOVA did not reveal any significant interaction.

Figure 4.5: Interaction of Modeling Notations with Expertise Co-factors



There is no interaction between scenario, session and modeling notation. There is a slight improvement in actual comprehension in favor of the risk model based on the Online Banking scenario as it is clearly more familiar than a Health Care Network. The improvement between two sessions is due to the learning effect, the participants became experienced in fulfilling comprehension task throughout the sessions.

Figure 4.6: Interaction of Scenario and Session vs Modeling Notation – Study 4.2

4.6 Discussion and Implications

In this section we discuss our results with respect to the hypotheses presented in Section 4.3.6. We also discuss possible explanation of the outcomes and their implications to research and practice.

The first null hypothesis $H4.1_0$ (about no difference between tabular and graphical risk models in the level of comprehensibility when performing comprehension task) *can be rejected for both precision and recall.* The second null hypothesis $H4.2_0$ (no difference between simple and complex questions in the level of comprehensibility when performing comprehension task) *can be rejected only for tabular representation*, but not for the graphical one.

The results show that, overall, the tabular risk model is more effective than the graphical one when stakeholders need to find relevant information about threats, vulnerabilities, or other elements of risk models. The participants who applied the tabular risk model gave more precise and complete answers to the comprehension questions. Regarding the perceived comprehensibility (Q5 and Q6 in post-task questionnaire) of the two risk modeling notations the participants showed equal preference for tabular and graphical risk models. These results are consistent across both our studies.

In this respect we argue that the difference in precision between the risk models could be explained by the cognitive fit theory itself if we do not unnecessarily restrict spatial relationships to graphs. Indeed, it can be seen that *tables also capture some aspects of linear spatial relationships.* In tabular risk models the name of the column identifies the type of a risk element (e.g., assets, threats, vulnerabilities, impact, likelihood, and security controls) and each row relates elements to each other. Hence, we can consider the proximity of cells along a row or along a column as a simple spatial relationship. Further, it is necessary to identify elements belonging to some classes, and tabular models make it easier to search for specific risk elements.

In contrast, locating and searching is less immediate in graphical risk models because in these models the risk elements are identified by means of graphical icons or positioning on the arrows between model elements that first must be learned by the participants in order to locate these elements. For complex comprehension tasks the lack of difference in precision between two risk models may be due to the fact that the task involves identifying complex spatial relationships, and graphical risk models provide a better overview of the system's risks that counterbalances the immediateness of tabular risk models. This theory could be tested by performing additional experiments in which significantly more complex questions are asked in order to determine whether there is a sweet spot where graphical models are easier to understand than tabular models. If the models were to get too large we assume that both tabular and graphical models would produce poor results.

Implications for practice Because tabular risk model was found to be more effective in extracting relevant information about security risks we recommend to adopt tabular representation when security risk assessment results have to be communicated to different stakeholders. In case of a wide range of stakeholders it is likely that some of them may not know a particular graphical risk modeling notation, while tables provides notation represented in a natural language. The stakeholders also may benefit from using "look-up" bonus of tables with filters and sorting option in the tables.

The importance of our study is that we investigated *a)* the effectiveness of tabular and graphical risk modeling notations in extracting correct information about security risks and *b)* the effect of task complexity on the level of comprehension of risk models by non-security experts.

4.7 Threats to Validity

In this section we discuss the main threats to validity.

Construct validity threats are mainly due to the method used to assess the outcomes of tasks. In our experiments the main threat to construct validity is related to the design of the questionnaires to assess the comprehension level of the participants and the risk models. To eliminate any potential bias introduced by a particular researcher, the questions and the risk models were checked by five researchers independently. The post-task questionnaire was designed based on previous studies [37, 91]. However, the design of the questionnaire may be strongly favoring one treatment over the other. Inspired by similar studies [41, 108], we used the names of element types in the question statements.

This may work in favor of the tabular risk model as the graphical model is more difficult to navigate and reply "look-up" questions. However, our data showed different. If we look at Figure 4.4a, in Study 1 the drop in precision of responses between simple and complex questions is very small for graphical representation and more evident for the tabular one and the difference in recall is similar to both representations. In Study 2 the drop in precision and recall is consistent for both representations. Also a significant part of the participants (39% in study 1 and 50% in study 2) used search in PDF documents with graphical risk models (see Tables 4.14a and 4.14b). An alternative way to validate whether the availability of textual labels has an effect on comprehensibility, is to compare tabular model with a UML-based graphical risk model containing names of element types as a part of representation.

Another threat can be cause by self-evaluation the level of knowledge in related areas (i.e. Security, Modeling, Domain Knowledge, etc.) that we collected with pre-task questionnaire. The source of threats in this case can be the so-called Dunning-Kruger effect [19], when less competent people tend to evaluate their knowledge too high suffering from internal illusion about their skills level, while highly competent people tend to

downgrade the level of their knowledge as they assume that others are more competent than themselves. We possibly observed this effect in the first study when the participants that evaluated themselves as "novices" in Modeling obtained better results than the "proficient" and "expert" participants who received worse results (see Figure 4.5a). However, this threat is not major to our study as we used self-evaluation of participants' knowledge only to control for possible effects, but not as the main factor or dependent variables.

Internal validity threats are mitigated by the use of randomized assignment to the treatments, even though some of the threats remain. The risk models used in the study are quite generic but were designed by real experts in CORAS and correspond to realistic models reporting risk assessment results. Also, the comprehension questions were validated by the risk model designers to ensure that the questions covered the comprehension of all risk modeling notation concepts. As can be seen from Tables 4.14a and 4.14b, most of the participants clearly understood the objectives of the study and the task that they had to perform.

Conclusion validity concerns the relationship between treatment and outcome. Aggregating data from different individual experiments may threaten validity due to the differences between the settings of the experiments and the groups of participants. However, we mitigated these threats by defining the family of experiments belonging to the same study (i.e., Study 1 or Study 2) as exact replications of the experimental procedure described in Section 4.3.7. Another threat to conclusion validity lies in the data analysis. We used a non-parametric test because it does not assume a normal distribution of the data. We used permutation test for two-way ANOVA only to find a possible interaction between the treatment and co-factors. The permutation test is a good alternative to standard test when the assumption about normal distribution is violated or the dataset is small [47].

External validity may be limited by the comprehension tasks and risk models used in the experiment and by the type of participants. Regarding the first point, we can say that the models chosen were created based on real application scenarios provided to us by an industrial partner. The HCN scenario was provided by IBM. Regarding the second point, others studies [122] have shown that students have a good understanding of the way that industry behaves, and may work well as participants in empirical studies. Moreover, students are not security experts and security standards place a big emphasis on "communicating risk", so that risk models/recommendations can be "consumed" by non-experts in security ([114, Section 2.1] or [5, Sec. 4.3]). Further studies may confirm whether or not our results can be generalized to more experienced participants (e.g., risk analysts and security professionals) and/or additional stakeholders' types who may be potential consumers of risk models (e.g., decision-makes or managers).

Chapter 5

Conclusions and Future Work

This chapter summarizes contribution of the thesis in relation to the research aim and research questions discussed in Chapter 1. Moreover, we discuss the limitations and future research directions in regard to the findings of the thesis.

5.1 Summary of the Thesis

This section summarizes the contribution of the thesis in relation to the research questions discussed in Chapter 1. We addressed the research questions as follows:

RQ1. Which SRA methods are actually effective?

We addressed this research question by conducting a pilot experiment and a series of three controlled experiments. We presented the results of the experiments in Chapter 2. In the first experiment with MSc students and IT Audit professionals we compared three classes of academic methods: threat-, goal-, and problem-based methods. The results showed that threat-based methods have higher overall perception and perceived ease of use than goal-based and problem-based methods. This can be explained by the fact that threat-based methods have a clear process helping participants to identify threats and security controls and use a graphical notation to present the results. The first experiment also revealed no difference in the perceived usefulness of the different classes of methods.

In the next two experiments with MSc students we focused on the comparison of visual and textual methods. The results showed that there is no clear difference between textual and visual methods with respect to their actual effectiveness, while participants reported higher preference for the visual methods over the textual one with respect to perceived easy of use, perceived usefulness, and intention to use.

RQ2. How can we validate and compare different SRA methods?

To address this research question, in Chapter 2 (Figure 2.4) we propose an evaluation framework that defines a formal procedure for comparison of SRA methods and evaluation of results of methods application with help of external industrial experts. We use this

framework to conduct empirical studies in regard to the research questions $RQ1$ and $RQ3$. We believe that the availability of this framework would allow researchers to conduct empirical evaluations of proposed methods and compare their results with the results of other experiments conducted based on our framework.

RQ3. What criteria define the success of an SRA method?

To address this research question we built a theoretical model (Figure 2.10 in Chapter 2) that extends MEM and hypotheses different features of SRA methods which determine the methods' actual and perceived efficacy. The model is based on the qualitative and quantitative results of three empirical studies. The qualitative analysis revealed that the main drivers for method's perceived ease of use is i) the presence in the method of clear process that supports main steps of SRA like identification of assets, threats and security controls, and ii) availability of visual summary providing a global overview of SRA. If the visual notation does not scale well it may harm method's perceived usefulness. Perceived usefulness and actual effectiveness of the method can also be increased by providing to the analyst the security catalogues which helps in identification of threat and security controls.

Further, we investigated two relations suggested by the theoretical model. First, in Chapter 3 we studied the role of catalogues of threats and security controls in an SRA. The quantitative analysis showed that for novices in both domain and security expertise there is no difference between generic and domain specific catalogues. In contrast, professionals with domain expertise who applied the method with catalogues identified threats and controls of the same quality as security experts without catalogues. The qualitative analysis indicated that security experts have different expectations from catalogues than non-experts. Non-experts are mostly worried about the difficulty of navigating through the catalogue (the larger and less specific the worse it was) while experts found it mostly useful to get a common terminology and a checklist that nothing was forgotten. To summarize our findings, we proposed a theory to explain how different catalogues' features contribute into an effective risk assessment process for novices and experts in either domain or security knowledge.

Second, in Chapter 4 we explored the relation between risk modeling representation and method's perceived efficacy proposed by our theoretical model. We conducted a series of controlled experiments to answer the question: *"how comprehensible are different representation approaches for risk models?"* The results showed that tabular risk models are more effective than the graphical ones with respect to simple comprehension task and slightly more effective for complex comprehension task. We believe that these results can be explained by a simple extension of Vessey's cognitive fit theory [125] as some linear spatial relationships could be also (and possibly more easily) captured by tabular models. While both tabular and graphical risk models equally good support the complex comprehension task because the easiness of searching elements and relationships of tabular risk

models is compensated by the easiness of understanding the overall risk picture provided by graphical risk model.

5.2 Limitations and Future work

In what follows, we discuss limitations and future research directions that may expand our work:

– **Students as participants.** The main limitation of our work is that most experiments involved MSc students and very few experiments were conducted with professionals. This is a common problem of controlled experiments in SE [110]. Thus, it important to validate the findings reported in this thesis with security professionals. This could give us better evidence that our findings are related to industrial practice.

– **SRA actual efficiency.** Another possible research line is an evaluation of the effort required to conduct SRA. This type of experiments requires very precise metrics and data collection approach allowing to quantify participants' efforts on SRA. Existing empirical studies employ mainly self-reporting (e.g. diary or self-estimation) or software-aided effort data collection approaches. The first approach is quite subjective and can be biased by not very motivated participants or participants that would like to impress others (e.g., students participating in the experiment which is a part of the study course). The second approach does not work well in case the object of the study requires activities not supported by computers, e.g., brainstorming or creative thinking. This challenge is still an open question for our topic.

– **Improve existing SRA methods.** A logical extension of the current work is the creation of a *mix-method incorporating the strong features* from both tabular (e.g., clear process, tabular summary, catalogues support) and graphical methods (e.g., graphical risk models), and providing a good tool support. However, this work require careful implementation and validation in order to avoid the creation of another hard to use method.

– **Automated SRA.** Having a tool support for SRA raises another question: *which research methods are appropriate for the evaluation of automated SRA methods?* The evaluation framework proposed in this thesis compares SRA methods in general and does not investigate such aspects as usability of methods' tool. We can propose to consider the use of techniques to study usability from Human-Computer Interaction domain [113]. Also it is important to *find a balance between human expertise and automation.* This requires an analysis of currently used SRA techniques and best practices in order to evaluate what can be successfully automated using existing technologies.

– **Grounded theory validation.** In Chapter 3 we proposed a *theory for understanding* [34] the role of catalogues in SRA. The initial validation of the theory was done in two ways. First, we quantitatively evaluate the effect of catalogues on the quality of SRA by comparing results of non-security and security professionals who conducted SRA with catalogues vs. results of security professionals who performed SRA without catalogues. The main limitation of this part is a small number of participants. It would be interesting to replicate this study with more professionals to confirm findings of the experiments. Second, we conducted focus groups with domain professionals and collected their qualitative opinion about catalogue's features. This part would benefit from further studies aiming to quantitatively validate the relationships between the theoretical concepts of the model similar to the work by Sabi et al. [93].

– **Risk Models Comprehension and Memorization.** In Chapter 4 we investigated the effectiveness of graphical and tabular risk modeling notations in extracting correct information about security risks. The main limitations of this study are *a)* that it includes only two risk modeling notations and *b)* involved only students. Also our comprehension task is limited by a "look-up" nature which might favoring tabular representation. To validate this we need to extend our research questions and to investigate which representation better supports memorization of information presented in models, a suggested by one of our reviewers. The experiment can use existing comprehension task but the models will be provided to participants just for a limited time to read and memorize. Then participants have to answer questions without having the models available.

– **Comprehension Task Complexity.** A task complexity factor requires further investigation as well. Our results showed that tabular representation is prone to questions' complexity, while graphical representation seems to be equally good for both simple and complex questions. Therefore, task complexity should be always taken into account when researchers investigate the comprehensibility of different representations and not only in security.

Chapter 6

Detailed Experimental Data for Chapter 2

6.1 Experiment 2.1

The *goal* of the experiment was to compare different classes of methods to identify threats and security controls: threat-based, goal-based and problem-based methods. It involved all methods listed in Table 2.3.

6.1.1 Experiment Execution

Table 6.1 summarizes the different phases of the execution of the first experiment by reference to our protocol.

The experiment uses a between-groups design: each group applied one method to one of industrial application scenarios. Groups were randomly assigned to methods and scenarios.

During the experiment we have collected an average of 37 responses out of 40 for each questionnaire sampling. We discarded questionnaire responses of some participants due to incompleteness (i.e., they did not answer all questions in a questionnaire). We accepted in total 147 questionnaire responses that were correctly completed, i.e. 91,9% response rate. We collected 203 post-it notes with methods advantages and disadvantages during post-it notes sessions and 5 hours of audio of focus group interviews. Participants delivered 15 group reports.

Table 6.2 shows for each method the number of groups and participants who have applied it. Groups were randomly assigned to methods and scenarios.

6.1.2 Method Designers Evaluation

Table 6.3 reports the results of reports evaluation by method designers.

Table 6.1: Execution Details – Experiment 2.1

	Step	Date	Description
	Training		
Trento	$M1$	May 7, 2012	Distribution of background and demographics questionnaire.
	$E1$	May 9-10, 2012	Participants attended lectures on the application scenarios.
	$E2$	May 8, 2012	Participants attended lectures about the methods.
	$M2$	May 10, 2012	Distribution of post-training questionnaire.
	Application		
	$E3$	May 10-11, 2012	During the first application phase groups spent 8,5 hours working on risks, and security and privacy requirements for each method.
	$E4$	May 11, 2012	Groups gave a 10 minutes presentation of preliminary results.
	$M3$	May 11, 2012	Distribution of post-task questionnaire about participants' perception of the method after the first application session.
Paris	$E3_2$	June 14, 2012	Groups had 4 hours to finalize the security and privacy analysis using the assigned method.
	$E4_2$	June 15, 2012	Groups gave a 10 minutes presentation of the results of their analysis using the assigned method.
	$M3_2$	June 15, 2012	Distribution of post-task questionnaire about participants' perception of the method after the first application session.
	Evaluation		
	$M4$	June 15, 2012	Participants took part in focus groups interviews and post-it notes session.
	$E5$	June 30, 2012	Groups delivered final reports.
	$M5$	June 30, 2012	Distribution of post-task questionnaire about quality of experiment organization.
	$M6$, $M7$	July 1-15, 2012	Method designers and domain experts assessed groups final reports.

Table 6.2: Experimental Design – Experiment 2.1

Method/Scenario	Smart Grid	eHealth	No of Participants
CORAS	1	2	8
SREP	2	1	8
LINDDUN	1	2	7
SECURE TROPOS	2	1	9
SECURITY ARGUMENTATION	1	2	9
N. of Groups	7	8	42

6.1.3 Quantitative analysis

In this section, we report the results of post-task questionnaires analysis.

Questionnaire Analysis: As we have 5 methods and we want to find the difference in participants' perception of methods we can use the KW test. This test requires homogeneity of variances and observations independence. By design, participants' answers about the evaluated methods are totally independent of each other, i.e. we have independence within samples and mutual independence between samples. We test the homogeneity of variances with the Levene's test. The test returned p-value > 0.05 for all samples and, thus, the assumption on homogeneity of variances is met. Since the assumptions are satisfied, we can apply the KW test. We also conducted a post-hoc test to compare participants' answers for each pair of methods with the MW test with Holm correction [46, Chap. 14.2]. Besides conditions *a*) homogeneity of variances and *b*) observations independence the test requires that response variables are measured in ordinal scale but this assumption holds because our data are answers on a 5-point Likert scale. Therefore, we can conclude that MW is a suitable test.

Table 6.3: Reports Assessment by Methods Designers – Experiment 2.1

Group	Method	Method Completeness
Group 1	CORAS	3
Group 2	CORAS	4
Group 3	CORAS	2
Group 4	SECTRO	2
Group 5	SECTRO	3
Group 6	SECTRO	4
Group 7	SECARG	4
Group 8	SECARG	3
Group 9	SECARG	3
Group 10	SREP	2
Group 11	SREP	4
Group 12	SREP	3
Group 13	LINDDUN	4
Group 14	LINDDUN	2
Group 15	LINDDUN	4

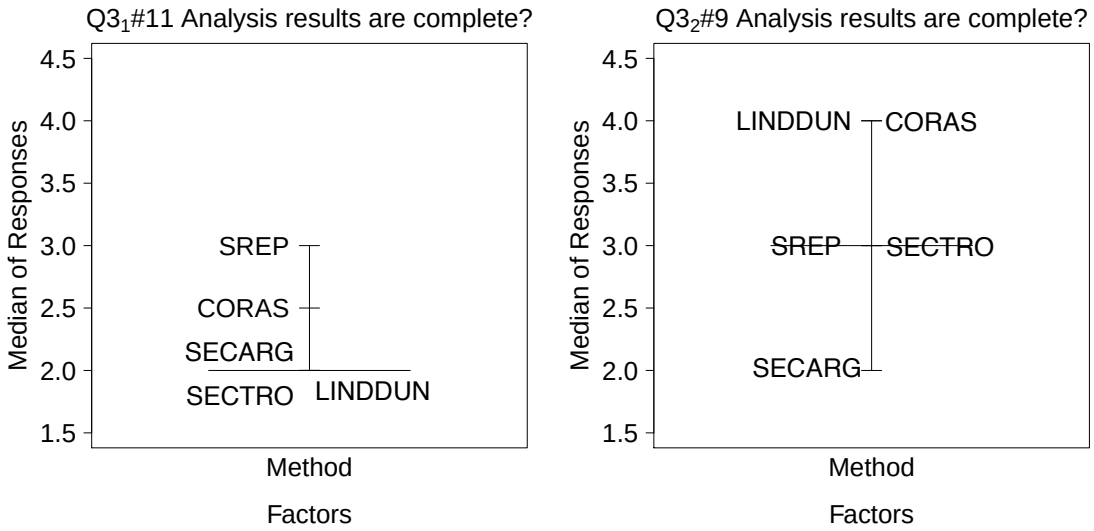
Table 6.4: Median Statistics and Results of the KW Test for Participants' Answers – Experiment 2.1

Question	Type	CORAS	SREP	LINDDUN	SECTRO	SECARG	Stat. sign.
BEFORE APPLICATION (One day after the Training)							
$Q3_1$#5: Overall applying the method was?	PEOU	3	4	3	2	3	
$Q3_1$#8: Conceptual model was easy to understand?	PEOU	3	4.5	4	2	3	*
$Q3_1$#11: Results of your analysis are complete?	PU	2.5	3	2	2	2	•
Overall $Q3_1$	PE	3	4	3	2	3	**
AFTER APPLICATION (One month after the Training with at least 2 full days of application)							
$Q3_2$#5: Overall applying the method was?	PEOU	3	4	3	3	3	**
$Q3_2$#6: Conceptual model was easy to understand?	PEOU	4	4	4	3	3	**
$Q3_2$#9: Results of your analysis are complete?	PU	4	3	4	3	2	
Overall $Q3_2$	PE	4	4	3.5	3	3	***

• - p-value <0.1, * - p-value <0.05, ** - p-value <0.01, *** - p-value <0.0001

For each question related to PEOU and PU, Table 6.4 reports the median value of participants' answers given for each method at the beginning ($Q3_1$) and at the end ($Q3_2$) of the Application phase, and the level of statistical significance based on p-value returned by the KW test. The table also reports the median of participants' answers related to methods' overall perception (questions #5, #8 and #11 in $Q3_1$ and #5, #6 and #9 in $Q3_2$) with the p-value returned by the KW test.

The overall perception of SREP is higher than the one of other methods at the beginning of the Application phase. At the end of the Application phase also CORAS' overall perception is definitely higher than the one of other methods. We had similar results for PEOU: at the beginning of the Application phase SREP's PEOU is higher than the one of other methods, and at the end of the Application phase we see that also CORAS' has higher PEOU than other methods. In case of PU the KW test did not reveal any



(a) At the beginning of the Application phase

(b) At the end of the Application phase

Immediately after the training, only the textual threat-based method (SREP) is perceived as the most useful. After a calendar month of remote application and almost two days full time of controlled application, two others threat-based methods (CORAS and LINDDUN) are perceived as useful. The observation is significant with 10% level after the training and not significant after the application (KW test).

Figure 6.1: Responses to the Question about Completeness of Analysis Results – Experiment 2.1

statistically significant difference between the methods.

We have also analyzed the individual questions to understand whether the difference in methods' overall perception could be due to PEOU or PU.

Figure 2.7 in Chapter 2 shows the medians of participants' answers to question about understandability of conceptual model at the beginning ($Q3_1$ #8) and at the end ($Q3_2$ #6) of the Application phase. We can observe that participants' perception of the methods varies between the beginning and the end of the phase. SREP is better than other methods with respect to understandability of the conceptual model both at the beginning and at the end of the phase with statistical significance.

The results of the KW test run for the questions about difficulty of method application and about understandability of conceptual model reveal that SREP's and CORAS' perceived ease of use is higher than s other methods with statistical significance. Figure 6.1 represents the median of participants' answers about completeness of analysis results at the beginning ($Q3_1$ #11) and at the end of the Application phase ($Q3_2$ #9).

As for the question on understandability of conceptual model, participants' perception of the methods with respect to completeness of analysis results varies between the beginning and the end of the Application phase. At the beginning of the Application phase, participants' perceived usefulness of SREP is higher than the other methods. Instead, at the end the perceived usefulness of CORAS and LINDDUN is higher than the other

Table 6.5: Qualitative Coding of Participants' Statements – Experiment 2.1

(a) Positive and Negative Aspects Influencing PEOU

PEOU Category	CORAS	SREP	LINDDUN	SECTRO	SECARG	Total
Positive Aspects						
Clear Process	16	9	9	7	9	50
Easy to Understand	3	15	3	4	4	29
Easy to Use	5	17	10	3	6	41
Formal Language	1			1	8	10
Visual Summary	15		9	4	1	29
Total Pos PEOU	40	41	31	19	28	159
Negative Aspects						
No clear process	6	5		7	10	28
Not easy to understand	2	5	3	5	6	21
Not easy to use	4	6	3	5	11	29
Primitive tool	9			11	5	25
Redundant Steps	6	3	1			10
Too time consuming	4	7	7	4	3	25
Total Neg PEOU	31	26	14	32	35	138
Total PEOU	71	67	45	51	63	297

(b) Positive and Negative Aspects Influencing PU

PU Category	CORAS	SREP	LINDDUN	SECTRO	SECARG	Total
Positive Aspects						
Help to Model	4			7	4	15
Help in Identifying Sec./Privacy Req.	1		1	2	6	10
Security/Privacy Specific		8	4		1	13
Total Pos PU	5	8	5	9	11	38
Negative Aspects						
Theoretical	1	6	7		4	18
Total Neg PU	1	6	7	0	4	18
Total PU	6	14	12	9	15	56

methods.

6.1.4 Qualitative analysis

Tables 6.5a and 6.5b present the coding results of questionnaires' open questions, post-it notes and focus group interviews transcripts related to PEOU and PU, respectively.

Each table presents for each evaluated method the categories that have a positive or negative impact on methods' PEOU and PU, and the total number of statements made by the participants for each of them. The number of statements is used as a relative indicator of category's importance.

In what follows we summarize the main aspects that may influence methods' PEOU and PU.

Perceived Ease of Use: Below we discuss the top PEOU categories. Participants made a significant number of statements (25% of positive and 20% of negative statements) supporting the importance of methods' ease of use.

Clear process has been reported as the main aspect that could affect PEOU of a method (31% of positive and 21% of negative statements). For CORAS (40% of positive statements) and LINDDUN (29% of positive statements and no negative) having a clear process positively affects their PEOU, while for SECURE TROPOS and SECURITY ARGUMENTATION there is no clear consensus among participants. Here are examples of statements made by participants about process of these methods: "For me it was very clear steps from the first till the last one." (CORAS); "The process is very clear and it is easy to understand the method." (LINDDUN); "Clear identification of security requirements and goals." (SECURITY ARGUMENTATION); "I think the process of the method is heavy, slow, complex to follow." (SECURITY ARGUMENTATION).

Easy to understand is another category that has an impact on participants' PEOU. Participants made 18% of positive and 15% of negative statements about it. For example, 37% of positive statements made about SREP were related to this aspect, while there is no consensus about ease of understanding of SECURE TROPOS (18% of positive and 16% of negative statements) and SECURITY ARGUMENTATION (13% of positive and 17% of negative statements). Examples of statements made by participants related to this category are the following: "The process is not so technical, so it is easy to understand." (SREP); "Some constructs are not the same with the other tools, which makes understanding of the concept difficult." (SECURE TROPOS).

Visual summary has positive impact on method PEOU (18% of positive statements). In CORAS are 37.5% of positive statements and in LINDDUN are 29% of positive statements about this aspect. Here are some examples of participants' statements about it: "The explicit description of the analysis process with diagrams." (CORAS); "Data flow diagram based. Method is clear and easy to follow and is focused on data flow diagrams." (LINDDUN).

Perceived Usefulness: Here we discuss the most emerged PU categories.

Help to model is a category which has a great impact on methods' PU (24% of positive statements). SECURE TROPOS has 54% of positive statements are about it. Here is an example of statement made by participants about this aspect: "I liked the fact that it helps you to model the use case that you are treating."

Security/privacy specific is another important PU category (25% of positive statements). For example, SREP is security specific method because 73% of positive statements are about it or LINDDUN is as privacy specific methods because 50% of positive statements are about it. Participants made the following statements related to this category: "The thing I found positive in the process is the fact that I had to do some research on the security aspects." (SREP); "Focus on usually neglected aspects of privacy." (LINDDUN).

Table 6.6: Execution Details – Experiment 2.2

Step	Date	Description
Training		
$E1$	October 2012	Participants attended lectures on the application scenario (2 hours).
$E2$	October 2012	Participants attended lectures about the methods (2 hours for each method).
$M1$	October 2012	Distribution of background and demographics questionnaire.
Participants were divided in groups and received their assignments.		
Application session was repeated for each task, i.e. we had four application sessions.		
	2012	Participants attended lecture on the threats and possible security controls specific for the task but not concretely applied to the scenario (2 hours).
$E3_1$	2012	Groups had one week to identify threats and security controls specific for the task using the assigned method.
$E4_1$	2012	Groups gave a 10 minutes presentation of preliminary results of the methods application and received feedback from the trainer.
$E3_1$	2012	Groups had one week to deliver an intermediate report to get feedback.
Evaluation		
$E5$	January 2013	Groups delivered final reports.
$E4$	Mid January, 2013	Groups gave a presentation summarizing their work in front of the experimenters and of the domain expert.
$M6$	Mid January, 2013	Domain expert evaluated the quality of the threats and the security controls proposed by groups.
$M3$	January 2013	Distribution of online post-task questionnaire about participants' perception of the methods.
$M4$	January, 2013	Participants shared their perception of the methods during the individual interview with one of the experimenters.

Table 6.7: Experimental Design – Experiment 2.2

Scenario	Method	
	Visual	Textual
Mgmnt	6	10
WebApp/DB	9	7
Net/Teleco	9	7
Mobile	8	8

6.2 Experiment 2.2

The *goal* of the second experiment was to evaluate and compare the two best methods emerging from the first experiment with respect to their *effectiveness* in identifying threats and security controls, and participants' *perception* of the two methods.

6.2.1 Experiment Execution

Table 6.6 shows the timeline and details of the execution of the second experiment.

To ensure a sufficient number of observations to produce significant conclusions we chose a within-subjects design where all groups apply both methods. To avoid learning effects groups identified threats and controls for four different tasks of the Smart Grid application scenario. Each group applied the visual method (CORAS) to exactly two different tasks and the textual method (SREP) to the remaining two tasks. For each task, the method to be applied by groups was randomly determined. Table 6.7 shows for each task the number of groups assigned to visual and textual methods.

As presented in Table 6.6 at the step $M4$ we conducted individual interviews with participants to collect their perception of methods. All interview were conducted according to interview guide which contained open questions about the overall opinion about methods, their advantages and disadvantages, the difficulties encountered during the application of methods and the main differences among them. The interview questions were the same for all interviewees even though some specific questions were added for some of participants when their answers to the post-task questionnaire were contradictory. The questions are reported in Section 6.5.

The questionnaire distributed at the step $M3$ was adapted from the questionnaire reported in [86] which was inspired to TAM [14]. The questionnaire consisted of 22 questions which were formulated in an opposite statements (positive statement on the right and negative statement on the left) format with answers on a 5-point Likert scale. To avoid that participants answered on "auto-pilot", some of the questions (e.g. Q2, Q10, Q13) were given with the most positive response on the left and the most negative on the right. The questionnaire is reported in Table 6.17 in Section 6.4.

6.2.2 Quantitative Analysis

This section presents the results of quantitative analysis of reports and post-task questionnaire.

Report Analysis: To assess the effectiveness of visual and textual methods, final reports delivered by groups were coded by researchers to count the number of threats and security controls. The groups who have got from a domain expert at least one assessment higher than *Generic* for threats or security controls were classified as *Good Groups*.

The experimental design of our second experiment is two-factor (the method and the task). Thus, we can use the two-way ANOVA or its non-parametric analog, the Friedman test, to analyze the number of threats and security controls identified with each of two methods and for each of four tasks. We have observation independence by design because groups worked individually. This gave us independence within sample and mutual independence within sample as the tasks were different. We applied the Levene's test to evaluate the homogeneity of variances. The test returned $p = 0.56$ for security controls, and $= 0.65$ for threats. So we have no evidence to reject this assumption. We verified whether dependent variables were normally distributed with the Shapiro-Wilk test. It returned $p = 7.7 * 10^{-3}$ for security controls and $p = 0.51$ for threats. Therefore, this assumption is met only for threats.

Since all ANOVA assumptions are satisfied for threats, we applied it to test the effectiveness of visual and textual methods with respect to the number of identified threats. For the security controls we used the Friedman test.

The results of reports analysis show that the visual method is more effective in identify-

Table 6.8: ANOVA for Threats – Experiment 2.2

(a) All Threats

Variable	DF	Sum Sq	Mean Sq	F-value	P-value
Group	14	439.0	31.36	2.047	0.040164 *
Method	1	260.4	260.42	17.003	0.000195 ***
Task	3	431.1	143.70	9.383	9.01e-05 ***
Method:Task	3	12.7	4.24	0.277	0.841597
Residuals	38	582.0	15.32		

(b) Good Threats

Variable	DF	Sum Sq	Mean Sq	F-value	P-value
Group	5	257.00	51.40	3.812	0.030034 *
Method	1	294.00	294.00	21.805	0.000683 ***
Task	3	142.87	47.62	3.532	0.051995 •
Method:Task	3	18.32	6.11	0.453	0.720511
Residuals	11	148.31	13.48		

• - p value <0.1, * - p <0.05, ** - p <0.01, *** - p <0.001

Table 6.9: Friedman test for Threats – Experiment 2.2

Variable	Chi-square	DF	F-value	P-value
Method (all threats)	5.4	1	7.875	0.0089 **
Task (all threats)	12.66	3	5.48	0.0029 **
Method (good threats)	15	1	Inf	0.0000 ***
Task (good threats)	8.71	3	3.36	0.0469 *

• - p value <0.1, * - p <0.05, ** - p <0.01, *** - p <0.001

Table 6.10: Friedman test for Security Controls – Experiment 2.2

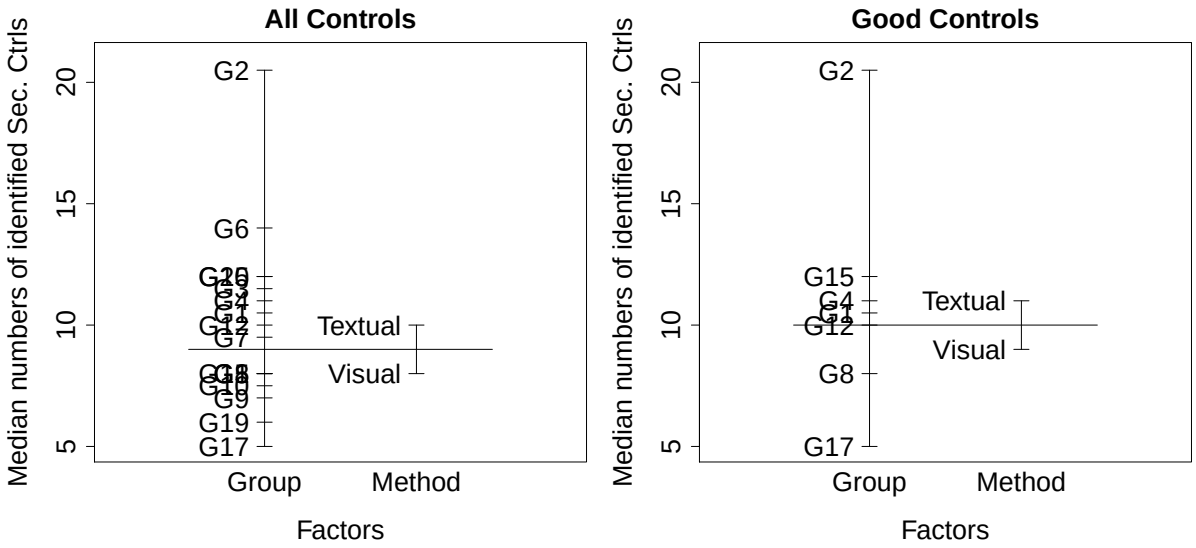
Variable	Chi-square	DF	F-value	P-value
Method (all controls)	0.692	1	0.68	0.4172
Task (all controls)	7.152	3	2.65	0.0614 •
Method (good controls)	9	1	19.29	0.0071 **
Task (good controls)	27.85	3	22.73	0.0145 *

• - p value <0.1, * - p <0.05, ** - p <0.01, *** - p <0.001

ing threats than the textual one. This result is confirmed for all groups (Figure 2.8 (left)) and good groups (Figure 2.8 (right)). The results of the ANOVA test (Table 6.8) show that the effect of the applied methods on the number of identified threats is statistically significant for all groups and good groups. The results of the Friedman test in Table 6.9 show that this is statistically significant for both all groups and good groups.

However, as it shown in Figure 6.2, the textual method is slightly better than the visual one in identifying security controls and this is true for controls of any quality and good ones. The results of the Friedman test in Table 6.10 shows that the difference is not statistically significant for all controls, but it is statistically significant for specific controls.

Questionnaire Analysis: We have analyzed the responses to post-task questionnaire to determine if there is a difference in participants' perception of visual and textual



The textual method performed slightly better in security controls identification in both cases of all and good groups. However, the difference is statistically significant for good groups, but not for all groups (Friedman test)

Figure 6.2: Numbers of Identified Security Controls by Quality – Experiment 2.2

methods. All participants independently answered same post-task questionnaire about each method. Full questionnaire is reported in Section 6.4. We collected two paired samples of answers about both methods from our participants. Questions were formulated in an opposite statements format (positive statement on the right and negative statement on the left) with answers on a 5-point Likert scale. Therefore, to test the answers of participants we need to use the Wilcoxon test because responses are paired and have ordinal values. As our samples had ties, we have used the exact Wilcoxon signed-ranks test with Wilcoxon method for handling ties [9].

The analysis results are summarized and compared in Table 6.11. For each question, the table reports to which perception variable the question refers to (PEOU, PU, ITU), the median of answers by all and by good participants (the one who were part of groups that produced good quality threats and security controls based on experts' assessment), and the level of statistical significance based on the p-value returned by the Wilcoxon test. The table also reports the average of responses for each perception variable and for all questions related to perception (Q1-Q12).

6.2.3 Qualitative Analysis

For a better understanding of which features influence visual and textual methods effectiveness, we complemented our experiment by interviewing each participant for half an hour. Similar to the first experiment, the interview transcripts were analyzed by using coding technique. Table 6.12 presents the categories and the frequency of statements

Table 6.11: Wilcoxon Test of Responses of All and Good Participants – Experiment 2.2

The level of statistical significance is specified by • ($p<0.1$), or * (* $p<0.05$, ** $p<0.01$, *** $p<0.001$).

		All participants			Good participants		
Q	**Type**	**Median**		**Z**	**Median**		**Z**
		Textual	**Visual**		**Textual**	**Visual**	
With Group 2							
1	PEOU	3	4	-0.89	2.5	4	-2.14 *
2	PU	3	4	-2.07 *	3	4	-0.64
3	PEOU	3	3.5	0.35	4	4	-0.69
4	ITU	3	3	-0.23	3	4	-1.5
5	PU	3	3	0.25	3	4	-2.27 *
6	ITU	3	3	0.02	2.5	3	-1.26
7	PEOU	3	3	-0.22	3	3	0
8	PU	4	4	0.43	4	4	0
9	ITU	3	4	-0.94	3	4	-1.26
10	PEOU	4	4	-0.49	4	4	0.19
11	PU	3	3	-0.41	3	3.5	-1
12	ITU	3	4	-0.83	3	4	-1
13	Control	4	4	-0.19	4	4	-1.73
14	Control	4	4	3.13 ***	5	4	2.27 *
15	Control	4	4	0.9	5	4	1.34
16	Control	3	4	-2.49 *	3	4	-1.81
17	Control	3	4	-1.67	3	4	-2.33 *
PEOU		3	3	-0.74	3.5	4	-1.65 •
PU		3	4	-1.03	3	4	-1.97 •
ITU		3	3	-0.96	3	4	-2.46 *
Without Group 2							
1	PEOU	3	4	-0.9	2.5	4	-2.2 *
2	PU	3	4	-1.7	3.5	4	0.11
3	PEOU	3	3.5	0.18	4	4	-1.09
4	ITU	3	3	-0.23	3	4	-1.5
5	PU	3	3	0.39	3	4	-2.07 •
6	ITU	3	3	0.17	3	3	-1
7	PEOU	3	3	-0.64	3.5	3.5	-0.63
8	PU	4	4	1.11	4	4	1.34
9	ITU	3	4	-0.97	3	4	-1.41
10	PEOU	4	4	-0.49	4	4	0.19
11	PU	3	3	-0.41	3	3.5	-1
12	ITU	3	4	-0.95	3	4	-1.2
13	Control	4	4	0.29	4	4	-1.19
14	Control	4	4	3.05 **	4.5	4	2.12 •
15	Control	4	4	0.9	5	4	1.34
16	Control	3	4	-2.28 *	3	3.5	-1.51
17	Control	3	4	-1.34	3	4	-1.89
PEOU		3	3.5	-1.02	4	4	-2.13 *
PU		3	4	-0.46	3	4	-1.19
ITU		3	3	-0.97	3	4	-2.52 *

in each category made by the participants. We report here only categories for which participants made at least 10 statements.

Perceived Ease of Use: Here we discuss the aspects reported by participants related to PEOU with respect to findings of our first experiment (see Section 6.1.4). *Ease of use and remember* is very important aspect of methods' success and participants supported this fact (35% of positive statements made by participants related to this PEOU aspect). If in the first experiment participants thought that textual method (SREP) is better than visual method (CORAS) with respect to ease of use and understand, than in the second experiment they changed their opinion and reported that visual method is a "good

Table 6.12: Positive and Negative Aspects Influencing Method Perception – Experiment 2.2

(a) PEOU

PEOU Category	Vis.	Text.	Total
Positive Aspects			
Clear process	12	16	28
Easy to use and remember	17	11	28
Support visual summary	24	0	24
Total Pos PEOU	53	27	80
Negative Aspects			
No clear process	2	11	13
Primitive tool	20	0	20
Too time consuming	11	9	20
Total Neg PEOU	33	20	53
Total PEOU	86	47	133

(b) PU

PU Category	Vis.	Text.	Total
Positive Aspects			
Help to understand interdependencies	6	7	13
Help in brainstorming threats	21	15	36
Help in brainstorming sec. ctrls	12	24	36
Total Pos PU	39	46	85
Negative Aspects			
No help in brainstorming sec. ctrls	9	1	10
Visual summary does not scale	10	0	10
Total Neg PU	19	1	20
Total PU	58	47	105

methodology, not difficult to use. It is much clear to understand the security case there".

Clear process is one of the main aspects related to PEOU reported by participants. As in our first experiment, participants think that visual method has a clear process (23% of positive statements on visual method's PEOU), and they have no consensus about clear process of the textual method. Participants made the following statements about clear process of methods: "Well defined steps. Clear process to follow" (SREP).

Primitive tool. This category demonstrated greater impact on method's PEOU in the second experiment: 23% of all negative statements was made about CORAS tool. The major problems reported were bad memory usage that makes the tool too slow and the modeling feature of the tool that does not provide automatic support for diagrams generation (e.g. generating a treatment diagram from a threat diagram). Examples of typical statements for this category were: "The tool is not difficult to use but it is very slow. It is impossible to copy a diagram from a type of diagram to another. Objects have no references between the diagrams. Changes on an object in a diagram are not reflected on the same object in other diagrams." and "The tool takes too much to arrange things. Drawing assets and threats is not easy. When the diagrams are too large, the tool occupies too much memory".

Visual summary. There is no surprise in this PEOU category. Like in the first experiment we observed the strong positive impact (45% of positive statements on visual method's PEOU) of visual summary on methods' perception. Indeed, diagrams give an overview of assets, possible threats scenarios and treatments. A typical statement made by participants referring to this advantage was: "Diagrams are useful. You have an overview of the possible threat scenarios and you can find links among the scenarios".

Perceived Usefulness: Here we discuss aspects reported by participants related to PU with respect to findings of our first experiment.

Help in identifying threats and *Help in identifying security controls.* In the second experiment participants classified visual method as helpful in identification of threats (54%

of positive statements on visual method's PU): "Yes it helped to identify which are the threats. In CORAS method everything is visualized. The diagrams helped brainstorming on threats." While the textual method according to participants is helpful in identification of security controls (53% of positive statements on textual method's PU): "SREP helped in brainstorming. The steps were pretty much defined. Step by step helped to discover more" and "SREP helped in brainstorming. The order of the steps helped to identify security requirements".

Visual summary does not scale. Participants of the second experiment also admitted that visual notation does not scale well for complex scenarios (53% of negative statements on visual method's PU). Typical statements in this category were: "The diagrams are not scalable when there are too many links" and "For big systems the diagrams would be very large. Even with the support of the computer it would be difficult to see them".

6.3 Experiment 2.3

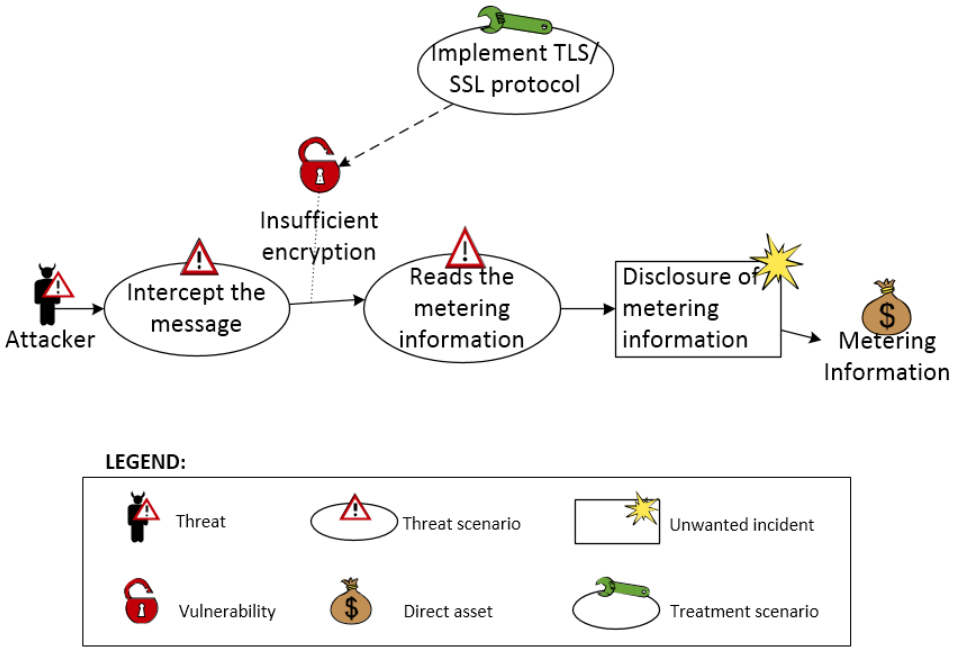
The *goal* of the third experiment was to generalize the previous results and investigated different textual method. As instance of textual method we selected SecRAM [21], an industrial method used by EUROCONTROL to conduct SRA in the air traffic management domain (ATM). SecRAM supports SRA process for a project initiated by an air navigation service provider, or ATM project, system or facility. It provides a systematic approach to conduct SRA which consists of five main steps: defining the scope of the system, assessing the impact of a successful attack, estimating the likelihood of a successful attack, assessing the security risk to the organization or project, and defining and agreeing a set of management options. As shown in Figure 6.3b) tables are used to represent the results of each step's execution in contrast to the CORAS which uses diagrams for this purpose (see Figure 6.3a).

6.3.1 Experiment Execution

Table 6.13 shows the timeline and details of the execution of the second experiment.

To ensure a sufficient number of observations to produce significant conclusions we chose a within-subjects design where all participants apply both methods. To avoid learning effects participants identified threats and controls for the two different tasks of the Smart Grid application scenario. Participants were randomly assigned to treatments: one half of participants applied first the visual method to Network Security task and then the textual method for the Web Application and Database Security task, while the other half applied methods in the opposite order. Table 6.14 shows for each task the number of participants assigned to visual and textual methods.

The post-task questionnaire distributed at the step $M3$ was revised and extended



(a) CORAS - Threat Diagram

Threat Agent	**Asset Attacked**	**Attack Likelihood**	**Justification**
Compromised MPO	SM, EMS	Probable	Can use message replay attack and access customer data.
Malicious attacker	EMS, HAN, SA, S&C	Probable	By Eavesdropping and Sniffing on the HAN, can use DoS attack to deny availability of HAN, Hacking the EMS and tampering the S&C and accessing the SA.

(b) SECRAM - Threat Agent Table

Figure 6.3: Examples of Visual (CORAS) and Textual (SecRAM) Methods' Artifacts Generated by Participants

version of the questionnaire from the second experiment. The questionnaire consisted of 31 questions which were formulated in an opposite statements (positive statement on the right and negative statement on the left) format with answers on a 5-point Likert scale. To prevent participants from "auto-pilot" answering, 15 out of 31 questions were given with the most positive response on the left and the most negative on the right. The questionnaire is reported in Table 6.18 in Section 6.4.

Similar to the second experiment, at the step $M4$ we conducted individual interviews with participants to collect their perception of the methods. The questions are reported in Section 6.5.

Table 6.13: Execution Details – Experiment 2.3

Step	Date	Description
Training		
$E1$	October 2013	Participants attended lectures on the application scenario (2 hours).
$E2$	October 2013	Participants attended lectures about the methods (2 hours for each method).
$M1$	October 2013	Distribution of background and demographics questionnaire.
Participants received individual assignments.		
Application session was repeated for each task, i.e. we had two application sessions.		
	2013	Participants attended lecture on the threats and possible security controls specific for the task but not concretely applied to the scenario (2 hours).
$E3_1$	2013	Groups had 2 weeks to identify threats and security controls specific for the task using the assigned method.
$E4_1$	2013	Groups gave a 10 minutes presentation of preliminary results of the methods application and received feedback from the trainer.
$E3_1$	2013	Groups had one week to deliver an intermediate report to get feedback.
Evaluation		
$E5$	January 2014	Groups delivered final reports.
$E4$	Mid January 2014	Groups gave a presentation summarizing their work in front of the experimenters and of the domain expert.
$M6$	Mid January 2014	Domain expert evaluated the quality of the threats and the security controls proposed by groups.
$M3$	January 2014	Distribution of online post-task questionnaire about participants' perception of the methods.
$M4$	January 2014	Participants shared their perception of the methods during the individual interview with one of the experimenters.

Table 6.14: Experimental Design – Experiment 2.3

Scenario	Method	
	Visual	Textual
Network	14	15
WebApp/DB	15	14

6.3.2 Quantitative Analysis

Report Analysis: Since a method is effective based not only on the quantity of results, but also on their quality, we asked two domain experts to independently evaluate each individual report. To evaluate the quality of threats and security controls experts used a four item scale: *Unclear* (1), *Generic* (2), *Specific* (3) and *Valuable* (4). In terms of the final assessment we observed that:

- the experts marked bad participants the same way,
- they consistently marked moderately good participants,
- a couple of participants were border line. In other words their threats and controls were neither definitely good nor bad.
- they had a different evaluation only for 3 out of 29 participants. This may be explained by different expertise of the domain experts: more management and seniority of one expert, more operational and junior other expert.

In order to validate whether the difference in experts' evaluation is statistically significant we run the Wilcoxon paired test. The results show that there are no statistically significant differences in the evaluations of two experts ($p = 0.58$).

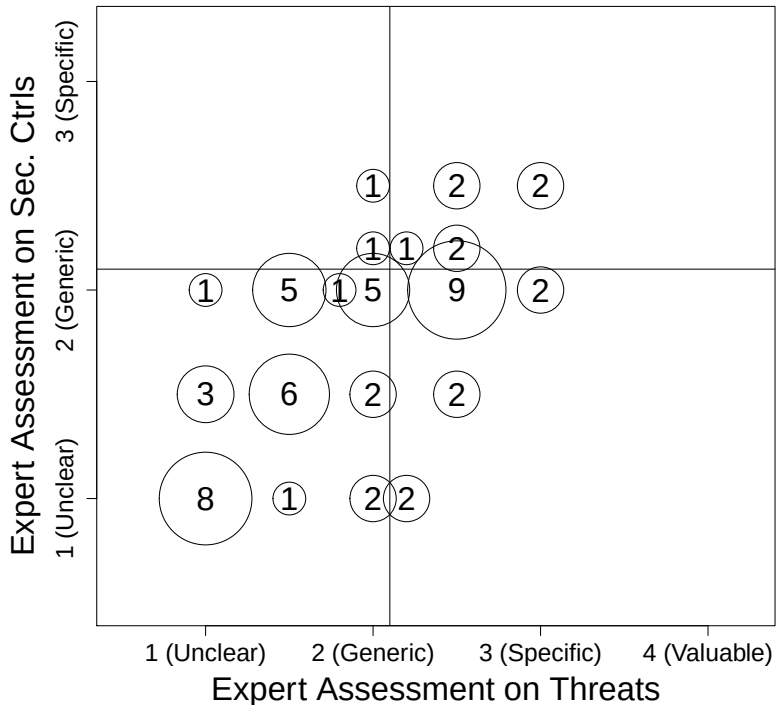


Figure 6.4: Overall Experts Assessment of Threats and Security Controls for the Two Tasks – Experiment 2.3

Figure 6.4 illustrates the average of the evaluation of two experts for all participants. As each participant applied one of the methods on both tasks, there are 58 method applications in total. The number inside each bubble denotes the number of methods' applications which achieved a given assessment for threats (reported on x-axis) and security controls (reported on y-axis). There were 24 method applications that generated some good threats and/or security controls. The remaining methods' applications delivered unclear and/or generic threats and security controls. We evaluated actual effectiveness of methods based on the number of *good* threats and security controls. In what follows, we will compare the results of all methods' applications with the results of those applications that produce good threats and security controls.

As the design of our experiment is two-factor (the method and the task), we could use the two-way ANOVA test or Friedman test (non-parametric analog of the ANOVA) to analyze the number of threats and security controls identified with each method and within each task. We have observation independence by design because participants' worked individually. This gave us independence within sample and mutual independence within sample, as tasks were different. We checked the homogeneity of variance with the Levene's test. This test returned p equal to 0.27 for threats and 0.52 for security controls. Therefore, we can assume homogeneity of variance for our samples. To check the distribution normality we used the Shapiro-Wilk normality test. This test returned p

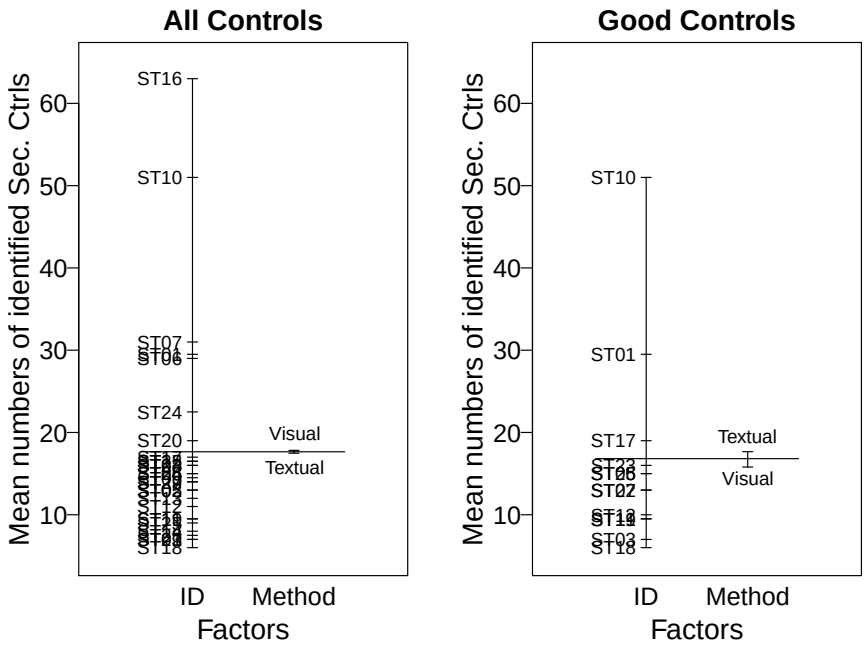


Figure 6.5: Numbers of Identified Security Controls by Quality – Experiment 2.3

= 0.01 for threats and 0.93 for security controls. So, normality assumption holds only for security controls.

Based on these results, we could use the Friedman test to analyze the difference in the number of threats and ANOVA test for security controls. However, since we also considered *good* results, we had unbalanced samples because some participants produced good threats and security controls for the application of one method while for the other method they did not. Therefore, we used the analog of the Friedman test, Skilling-Mack test [6], that can work with unbalanced samples for the analysis of difference in the *number of threats*, and the ANOVA test with Type II of Sum of Squares [60] for the analysis of difference in the number of *security controls*. The results related to threats identification are reported in Figure 2.9 in Section 2.7.

The results of reports analysis show that the textual method is more effective in identifying threats than the textual one for good participants only (Figure 2.9 (right)). But the results of the Skillings–Mack test did not confirm this (test returned $p = 0.17$).

Figure 6.5 shows that visual and textual methods produce the same number of security controls. This is attested also by the results of the statistical tests which showed there was no statistically significant difference in the number of security controls of any quality (Friedman test returned $p = 0.57$) and specific security controls (ANOVA test returned $p = 0.72$).

We also found that there is no statistically significance difference in the number of

threats and controls identified by participants within each security task.

Questionnaire Analysis: We compare answers of all participants with answers of those participants whose methods applications produced specific threats and/or security controls (denoted as *good participants* in what follows). We analyzed answers of all participants with the Wilcoxon test since the data are ordinal and responses of participants are paired. Instead, we used the MW test to analyze the answers of participants who produced specific results because some observations were unpaired. Since the MW test requires homogeneity of variance of samples, we checked this assumption. The Levene's test revealed that in general the variances of our samples are equal ($p = 0.95$). However, there is no equal variance for responses on overall PEOU of method ($p = 0.036$). Thus, we could not consider the results of the MW test of this category as valid.

Table 6.15 presents the results of questionnaire analysis. For each question, the table reports to which perception variable the question refers to (PEOU, PU, ITU), the mean of answers, and Z statistics returned by the test and the level of statistical significance based on the p-value returned by the test.

6.3.3 Qualitative Analysis

The interview transcripts were coded and analyzed using the list of core codes from the second experiment. Table 6.16 reports the positive and negative aspects of visual and textual methods that may affect PEOU and PU and *other* aspects that may influence methods' success. For each aspect we report the total number of statements made by participants as relative indicator of its importance. Here we report only the aspects for which participants made at least 10 statements.

Perceived Ease of Use: The main aspect influencing PEOU of visual method is that it provides a *visual summary* of the results of security analysis (29% of positive statements made by participants on visual method's PEOU). Examples of these statements are: "*there are many summary diagrams which are useful to summarize what has been done*" and "*the advantage is the visualization*". Another noteworthy positive aspect for visual method's PEOU is that the visual method has *clear process* (19% of positive statements):"*The advantages of CORAS is very clear structure*". Instead, the main aspects that can affect negatively the visual method's PEOU are that it is a *time consuming* method and it has a *primitive tool* (26% of negative statements). As participants indicated "*the diagrams are really time consuming*" and "*first I tried the CORAS tool. And somehow, it was confusing. So, I switched to the Visio*". Another negative aspect for visual method's PEOU is that the process has *redundant steps* (17% of negative statements): "*I think CORAS has some duplications.*".

The main positive aspect for the textual method's PEOU is *time effectiveness* (26% of positive statements): "*I used very little time to do my work*". Instead, there is no

Table 6.15: Mann-Whitney and Wilcoxon Tests of Responses of All and Good Participants – Experiment 2.3

The level of statistical significance is specified by • ($p<0.1$), or * (* $p<0.05$, ** $p<0.01$, *** $p<0.001$).

		All participants				Good participants		
Q	**Type**	**Mean**		Z_W	Z_{MW}	**Mean**		Z_{MW}
		Tex	Vis			Tex	Vis	
Q1	PU	4	4	-1.66 •	-1.63 •	4	4	-1.54
Q2	Control	4	4	0.07	-0.1	4	4	-0.41
Q3	Control	4	4	-0.37	-0.75	4	4	-0.75
Q4	PU	3	4	-2.37 *	-2.15 *	4	4	-1.34
Q5	PU	3	4	-2.03 *	-1.53	3	3	-0.47
Q6	PEOU	3	4	-2.7 *	-2.84 ***	3	4	-1.57
Q7	PEOU	3	4	-2.42 *	-2.5 *	3	4	-1.42
Q8	PU	4	4	-1.79 •	-1.69 •	4	4	-1.38
Q9	PEOU	2	4	-3.33 ***	-2.98 ***	2	4	-2.03 •
Q10	PU	3	4	-2.36 *	-2.63 *	3	4	-2.04 •
Q11	PU	3	4	-2.15 *	-1.89 •	3	3	-0.78
Q12	PU	3	4	0.61	0.09	3	3	-0.67
Q13	PU	3	3	0.41	0.06	3	3	-0.2
Q14	PU	4	4	-0.88	-0.73	4	4	-0.82
Q15	ITU	3	4	-1.83 •	-1.97 •	3	3	-0.55
Q16	ITU	4	4	-0.57	-0.66	4	4	-0.91
Q17	Control	3	4	-3.32 ***	-3.42 ***	2	4	-1.93 •
Q18	Control	3	4	-2.73 *	-2.66 *	3	4	-0.98
Q19	ITU	3	4	-2.26 *	-2.14 *	3	4	-1.85 •
Q20	ITU	3	4	-1.55	-1.39	3	4	-1.08
Q21	Control	4	3	1.95 •	1.89 •	3	3	-0.16
Q22	PU	3	3	-1.11	-0.89	3	3	-1.45
Q23	ITU	3	4	-1.52	-1.53	3	3	-0.95
Q24	ITU	3	4	-1.03	-1.05	3	3	-0.9
Q25	PU	3	4	-2.02 •	-1.63	3	4	-1.61
Q26	PU	3	4	-1.39	-1.47	3	4	-1.03
Q27	PEOU	3	4	-2.73 *	-2.78 *	3	4	-1.9 •
Q28	ITU	3	4	-1.19	-1.22	3	4	-1.31
Q29	ITU	3	4	-0.14	-0.39	3	3	-0.59
Q30	PEOU	3	4	-2.78 *	-1.91 •	2	3	-1.06
Q31	PEOU	3	4	-2.39 *	-2.07 *	2	4	-2.19 *
PEOU		3	4	-6.51 ***	-6.16 ***	2.5	4	-4.19 ***
PU		3	4	-4.82 ***	-4.56 ***	3	4	-3.88 ***
ITU		3	4	-3.57 ***	-3.67 ***	3	4	-2.94 ***

consensus among participants about other two aspects: *clear process* and *ease of use*. In fact, participants made a similar number of statements that indicate these aspects as both positive and negative: "*it's quite easy*" (positive statement) and "*it was sometimes a bit confusing how to apply the methodology*" (negative statement).

The main negative aspect (28% of negative statements) impacting textual method's PEOU is related to *poor worked examples* illustrating method application. As participants reported "*the main problem was about the example that it uses - instead of defining in more general way, and you are misguided by this example*".

Perceived Usefulness: There are two main aspects that could positively affect PU of visual method: *help in identifying threats* (55% of positive statements) and *security controls* (31% of positive statements): "*when you're doing a diagram you can actually see the flaw of the actions and it is easy to identify the threats, the attacks*" and "*I find it good*

Table 6.16: Positive and Negative Aspects Influencing Method Perception – Experiment 2.3

(a) PEOU

PEOU Category	Vis.	Text.	Total
Positive Aspects			
Clear Process	28	18	46
Visual summary	43		43
Time effective	7	16	23
Easy to Understand	18		18
Worked examples	12	4	16
Easy for Customer	13	2	15
Total Pos PEOU	121	40	161
Negative Aspects			
Time consuming	36	7	43
Unclear Process	4	28	32
Primitive Tool	30		30
Poor worked examples	2	27	29
Not easy to Use	6	18	24
Redundant Steps	19	4	23
No Evolution Support	15	2	17
Not easy to Understand	3	11	14
Total Neg PEOU	115	97	212
Total PEOU	236	137	373

(b) PU and Other

PU Category	Vis.	Text.	Total
Positive Aspects			
Help in Identifying Threats	39	18	57
Help in Identifying Security Controls	22	16	38
Help to Model	10	2	12
Total Pos PU	71	36	107
Negative Aspects			
No Help in Identifying Security Controls	9	16	25
No Tool Support		21	21
Visual Complexity	17		17
Total Neg PU	26	37	63
Total PU	97	73	170

Other Category	Vis.	Text.	Total
Positive Aspects			
Catalog of Sec. Controls	23	31	54
Catalog of Threats	30	29	59
Total Pos Other	53	60	113

for finding some security requirements and risk". The negative aspect for visual method PU is that visual notation does not scale well for complex scenarios (65% of negative statements): "*these diagrams are getting soon very huge and very complex*".

Similarly, the main positive aspect for textual method PU is that "*it has detailed steps and helps to identify assets, threat agents and management options*" (50% of positive statements). Instead, there is no consensus among participants about the textual method helping in the *identification of security controls*. In fact, they made equal number of positive and negative statements about this aspect. Here are examples of typical statements made by participants about it: "*After we already known that our system description, the vulnerabilities, the threat or agents is easy to identify the control.*" (positive statement) or " *I can't say that they allow you to find the threat, the security control, whatever you want. It's just a framework to help you.*" (negative statement).

The most significant negative aspect mentioned for textual method's PU is the fact there is no software supporting execution of the textual method's steps: "*It is needed because it would save half of the time if the table were generated automatically*" (57% of positive statements).

Other Relevant Aspects: In participants' interview we also identified other possible aspect influencing methods' success. Participants think that both methods would benefit from availability of catalogues of threats and security controls: "*I think that SecRAM could just employ some catalog by default.*".

6.4 Post-Task Questionnaires

Table 6.17: Post-Task Questionnaire – Experiment 2.2

Left statement	1 2 3 4 5	**Right statement**
Q1.I found X hard to use	❍ ❍ ❍ ❍ ❍	I found X easy to use
Q2. X made the security analysis easier than an ad hoc approach	❍ ❍ ❍ ❍ ❍	X made the security analysis harder than an ad hoc approach
Q3. X was difficult to master	❍ ❍ ❍ ❍ ❍	X was easy to master
Q4. If I need to identify threats and security requirements in a future project course, I would not use X	❍ ❍ ❍ ❍ ❍	If I need to identify threats and security requirements in a future project course I would use X
Q5. I would have found threats and security requirements more quickly using common sense	❍ ❍ ❍ ❍ ❍	X made me find threats and security requirements more quickly than using common sense
Q6. If I need to identify threats and security requirements in a future project at work, I would avoid X if possible	❍ ❍ ❍ ❍ ❍	If I need to identify threats and security requirements in a future project at work, I would use X if possible
Q7. I was often confused about how to apply X to the problem	❍ ❍ ❍ ❍ ❍	I was never confused about how to apply X to the problem
Q8. X made the search for threats and security requirements less systematic	❍ ❍ ❍ ❍ ❍	X made the search for threats and security requirements more systematic
Q9. If a company I'm employed by in the future discusses what technique to introduce for early security analysis argue and someone suggests X, I would be against that	❍ ❍ ❍ ❍ ❍	If a company I'm employed by in the future discusses what technique to introduce for early security analysis and someone suggests X, I would support that
Q10. X will be easy to remember (in case I must use it again in the future)	❍ ❍ ❍ ❍ ❍	X will be hard to remember (in case I must use it again in the future)
Q11. X made me less productive in finding threats and security requirements	❍ ❍ ❍ ❍ ❍	X made me more productive in finding threats and security requirements
Q12. If working as a freelance consultant for a customer who needs help finding security threats and security requirements to his software, I would not use X in discussions with that customer	❍ ❍ ❍ ❍ ❍	If working as a freelance consultant for a customer who needs help finding security threats and security requirements to his software, I would like to use X in discussions with that customer
Q13. X process is well detailed	❍ ❍ ❍ ❍ ❍	X process is not well detailed
Q14. A catalog of threats would have made the identification of threats easier with X	❍ ❍ ❍ ❍ ❍	A catalog of threats would have made the identification of threats harder with X
Q15. A catalog of security requirements would have made the identification of security requirements easier with X	❍ ❍ ❍ ❍ ❍	A catalog of security requirements would have made the identification of security requirements harder with X
Q16. X helped me in brainstorming on the threats for the tasks	❍ ❍ ❍ ❍ ❍	X did not help me in brainstorming on the threats for the tasks
Q17. X helped me in brainstorming on the security requirements for the tasks	❍ ❍ ❍ ❍ ❍	X did not help me from brainstorming on the security requirements for the tasks
Q18.* CORAS tool is hard to use	❍ ❍ ❍ ❍ ❍	CORAS tool is easy to use
Q19. The first task (Management) was very hard	❍ ❍ ❍ ❍ ❍	The first task (Management) was very easy
Q20. The second task (WebApp/DB) was very hard	❍ ❍ ❍ ❍ ❍	The second task (WebApp/DB) was very easy
Q21. The third task (Net/Telecom) was very hard	❍ ❍ ❍ ❍ ❍	The third task (Net/Telecom) was very easy
Q22. The fourth task (Mobile) was very hard	❍ ❍ ❍ ❍ ❍	The fourth task (Mobile) was very easy

* - This question is asked only in the questionnaire about CORAS

Table 6.18: Post-Task Questionnaire – Experiment 2.3 (Part 1)

Left statement	1 2 3 4 5	**Right statement**
Q1.The method defines the right level of granularity of asset, security risk and security control.	❍ ❍ ❍ ❍ ❍	The method defines the wrong level of granularity of asset, security risk and security control.
Q2. A catalog of threats would have made harder the identification of threats with this method.	❍ ❍ ❍ ❍ ❍	A catalog of threats would have made easier the identification of threats with this method.
Q3. A catalog of security controls would have made easier the identification of security controls with this method.	❍ ❍ ❍ ❍ ❍	A catalog of security controls would have made harder the identification of security controls with this method.
Q4. Overall, I think the method provide an effective solution to the identification of security risks	❍ ❍ ❍ ❍ ❍	Overall, I think the method does not provide an effective solution to the identification of security risks
Q5. Overall, I think the method does not provide an effective solution to the identification of security controls	❍ ❍ ❍ ❍ ❍	Overall, I think the method provides an effective solution to the identification of security controls
Q6. I found the method easy to adopt and use to different contexts.	❍ ❍ ❍ ❍ ❍	I found the method hard to adopt and use to different contexts.
Q7. Overall, I found this method difficult to use	❍ ❍ ❍ ❍ ❍	Overall, I found this method easy to use
Q8. Overall, I found this method to be useless	❍ ❍ ❍ ❍ ❍	Overall, I found this method to be useful
Q9. I found the procedure for applying the method complex and difficult to follow	❍ ❍ ❍ ❍ ❍	I found the procedure for applying the method simple and easy to follow
Q10. Using this method would make harder to identify security risks specific for the context.	❍ ❍ ❍ ❍ ❍	Using this method would make easier to identify security risks specific for the context.
Q11. Using this method would make easier to identify security controls specific for the context.	❍ ❍ ❍ ❍ ❍	Using this method would make harder to identify security controls specific for the context.
Q12. I believe that this method would reduce the effort required to identify security risks of complex systems	❍ ❍ ❍ ❍ ❍	I believe that this method would increase the effort required to identify security risks of complex systems
Q13. I believe that this method would increase the effort required to identify security controls of complex systems	❍ ❍ ❍ ❍ ❍	I believe that this method would reduce the effort required to identify security controls of complex systems
Q14. Using this method would make harder to find the most complete set of security risks.	❍ ❍ ❍ ❍ ❍	Using this method would make easier to find the most complete set of security risks.
Q15. If working as a freelance consultant for a customer who needs help in identification of security risks to his software, I would not use the method in discussions with that customer	❍ ❍ ❍ ❍ ❍	If working as a freelance consultant for a customer who needs help in identification of security risks to his software, I would use the method in discussions with that customer
Q16. If working as a freelance consultant for a customer who needs help in identification of security controls to his software, I would use the method in discussions with that customer	❍ ❍ ❍ ❍ ❍	If working as a freelance consultant for a customer who needs help in identification of security controls to his software, I would not use the method in discussions with that customer
Q17. Using this method would make harder to compare security risks identified in the risk assessment with security risks identified by other methods.	❍ ❍ ❍ ❍ ❍	Using this method would make easier to compare security risks identified in the risk assessment with security risks identified by other methods.
Q18. Using this method would make easier to compare security controls identified in the risk assessment with security controls identified by other methods.	❍ ❍ ❍ ❍ ❍	If a company I'm employed by in the future discusses what technique to introduce for identification of security risks and someone suggests the method, I would support that
Q19. If a company I'm employed by in the future discusses what technique to introduce for identification of security risks and someone suggests the method, I would be against that	❍ ❍ ❍ ❍ ❍	If a company I'm employed by in the future discusses what technique to introduce for identification of security risks and someone suggests the method, I would support that
Q20. If a company I'm employed by in the future discusses what technique to introduce for security controls identification and someone suggests the method, I would support that	❍ ❍ ❍ ❍ ❍	If a company I'm employed by in the future discusses what technique to introduce for identification of security controls and someone suggests the method, I would be against that
Q21. Using this method would make easier to update the risk assessment when something changes.	❍ ❍ ❍ ❍ ❍	Using this method would make harder to update the risk assessment when something changes.
Q22. The method makes the traceability of security controls down to security risk and assets harder.	❍ ❍ ❍ ❍ ❍	The method makes the traceability of security controls down to security risk and assets easier.
Q23. If I need to identify security risks in future project at work, I would avoid the method if possible	❍ ❍ ❍ ❍ ❍	If I need to identify security risks in a future project at work, I would use the method if possible
Q24. If I need to identify security controls in a future project at work, I would use the method if possible	❍ ❍ ❍ ❍ ❍	If I need to identify security controls in a future project at work, I would avoid the method if possible

Table 6.19: Post-Task Questionnaire – Experiment 2.3 (Part 2)

Left statement	1 2 3 4 5	**Right statement**
Q25. I believe that this method would make easier to find cost effective mitigation of security risks to an acceptable level	❍ ❍ ❍ ❍ ❍	I believe that this method would make harder to find cost effective mitigation of security risks to an acceptable level
Q26. The method made me more productive in identification of security risks	❍ ❍ ❍ ❍ ❍	The method made me less productive in identification of security risks
Q27. I found the method difficult to learn	❍ ❍ ❍ ❍ ❍	I found the method easy to learn
Q28. If I need to identify security risks in a future study project, I would use the method if possible	❍ ❍ ❍ ❍ ❍	If I need to identify security risks in a future study project, I would avoid the method if possible
Q29. If I need to identify security controls in a future study project, I would avoid the method if possible	❍ ❍ ❍ ❍ ❍	If I need to identify security controls in a future study project, I would use the method if possible
Q30. I am not confident that I am now competent to apply this method in practice	❍ ❍ ❍ ❍ ❍	I am confident that I am now competent to apply this method in practice
Q31. The method will be easy to remember (in case I must use it again in the future)	❍ ❍ ❍ ❍ ❍	The method will be hard to remember (in case I must use it again in the future)

6.5 Interview Guide

Table 6.20 presents the questions that were asked both for the visual (CORAS) and the textual method (SREP or SecRAM) and mapping of the questions to the experiments.

Table 6.20: Interview Guide

Q#	**Question statement**	**Exp. 2.2**	**Exp. 2.3**
1	What do you think about method?	×	×
2	Do you think the method is an easy method to apply? Why?		×
3	While applying the method where you got confused about how to apply it?	×	
4	Do you think the method helps you brainstorming? Why?	×	
5	Do you think the method helped you to identify threats?	×	×
5.1	Do you think the method helped you to identify threats in a reasonable time?		×
5.2	Do you think a catalog of threats would have made the identification easier? Did you use any catalog? Which one?		×
6	Do you think the method helped you to identify security controls?	×	×
6.1	Do you think the method helped you to identify security controls in a reasonable time?		×
6.2	Do you think a catalog of security controls would have made the identification easier? Did you use any catalog? Which one?		×
7	What are the advantages of the method?	×	×
8	What are the disadvantages of the method?	×	×
9	[CORAS only] Did you use CORAS tool? Which version of the CORAS tool did you use? If not, what did you use to draw the diagrams?	×	×
10	[SecRAM only] Do you think tool support is needed for SecRAM?		×
11	What do you think are significant differences between the two methods?	×	
12	Which task was according to you the most difficult? And why?	×	
13	Can you suggest possible improvements for the method?		×

Chapter 7

Additional Data for Chapter 3

7.1 Studies Using 5-item Likert Scale

In total our search returned 22 papers that were published in MISQ, 20 papers in ISR, and 14 papers in ManSci journals. After checking the papers, we obtained a sample of 7 papers that reported descriptive statistics (incl. number of participants, mean and standard deviation of each group) of dependent variables on a 5-item Likert scale. Table 7.1 describes the final set of selected papers.

Table 7.1: Studies Using 5-item Likert Scale

Authors	Title	Journal	Year	Dependent variables	#subj.
Ferratt et al.	Synergy and Its Limits in Managing Information Technology Professionals	ISR	2012	Job search behaviour	251
Gopal and Koka	The asymmetric benefits of relational flexibility: evidence from software development outsourcing	MISQ	2013	Requirements Uncertainty; Human Asset Specificity; Employee Turnover; Client MIS Experience; Quality	105
Gove and Parson	Is query reuse potentially harmful? Anchoring and adjustment in adapting existing database queries	ISR	2010	Reuse of queries results in higher confidence in query correctness	157
Maruping and Magni	Motivating Employees to Explore Collaboration Technology in Team Contexts	MISQ	2014	CT Exploration; Team Empowerment; Continued Intention to Explore; Continued Expectation to Explore; Perceived Usefulness; Facilitating Conditions; Personal Innovativeness with IT; Intention to Explore; Training; Task Interdependence	212
Montizaan et al.	The Impact of Negatively Reciprocal Inclinations on Worker Behavior: Evidence from a Retrenchment of Pension Rights	Man. Sci.	2016	Negative reciprocity; Positive reciprocity	5287
Phang et al.	What Motivates Contributors vs. Lurkers? An Investigation of Online Feedback Forums	ISR	2016	Civic skill	101
Sutanto et al.	Addressing the Personalization-Privacy Paradox: An Empirical Assessment From a Field Experiment on Smartphone Users	MISQ	2013	Excessive advertisements	60

Table 7.2: Results of the Coding Analysis for Each Focus Group and Overall

Codes – Success Criteria	**FG1**	**FG2**	**FG3**	**FG4**	**TOTAL**
Applicable to different domains	4	0	0	3	7
* Appropriate to the problem	4	1	1	0	6
Assessment techniques	0	0	0	0	0
Catalogue of threats and security controls	3	0	1	1	5
Clear steps in the process	8	1	3	4	16
Comparability of results	5	2	0	7	14
Compliance with standards	8	0	3	1	12
Comprehensibility of method outcomes	3	7	2	2	14
* Cost effective	3	0	1	0	4
Coverage of results	0	2	0	5	7
Documentation templates	5	2	1	0	8
* Domain expertise is needed	1	3	0	3	7
* Easy to apply	5	2	0	4	11
* Effective risk mitigation	6	4	3	3	16
Evolution support	1	3	0	1	5
* Granularity	0	1	7	2	10
Help to identify threats	2	1	2	2	7
Holistic process	0	1	0	2	3
* Interoperability	1	5	2	1	9
* Learn new things	1	4	1	1	7
Modelling support	1	0	0	0	1
Practical guidelines	11	0	1	2	14
* Record failed attacks	1	0	2	0	3
* Security expertise is needed	2	2	0	1	5
Specific controls	4	1	5	2	12
Systematic listing	0	0	0	0	0
Time effective	4	4	0	1	9
Tool support	3	1	4	4	12
* Traceability	1	2	1	0	4
Visualization	0	1	1	0	2
Well defined terminology	5	6	1	2	14
Worked examples	3	0	1	0	4
TOTAL	95	56	43	54	248

Categories – Tasks	**FG1**	**FG2**	**FG3**	**FG4**	**TOTAL**
Finding information	16	9	10	8	43
Presenting-sharing information	9	16	7	3	35
Validating information	12	5	3	9	29
TOTAL	37	30	20	20	107

Note: $*$ marks codes different from Table 6.16.

Table 7.3: Focus Groups Interview Guide

Q#	**Question statement**
1	What makes a security risk assessment methodology successful?
2	What are typical weaknesses of security risk assessment methodologies?
3	What factors influence your intention to use a methodology for security risk assessment?
4	Is compliance with your organizational requirements and procedures, an aspect that you consider when you select the security risk assessment methodology to use?
5	What makes a security risk assessment methodology easy to use?
6	What makes a security risk assessment methodology effective? (e.g in terms of identification of threats, security controls)

Figure 7.1: Quality Evaluation Guidelines for Experts

Report Quality Assessment by Domain Experts

Link to folder with reports:

Listed below are the criteria and marks followed by domain experts for the report quality assessment:

Scale	Threats Quality	Security Controls Quality
1 - Bad	Not clear which are the final threats for the scenario	Not clear which are the final security controls for the scenario
2 - Poor	Threats are present but are not specific for the scenario	Security controls are present but are not specific for the scenario
3 - Fair	Threats are present and SOME of them are related to the scenario	Security controls are present and SOME of them are related to the scenario
4 - Good	Threats are present and they are related to the scenario	Security controls are present and they are related to the scenario
5 - Excellent	Threats are present and they are major threats for the scenario	Security controls are present and propose real solutions for the scenario

On the page "Assessment Form" you will find the table with the following

Group ID	Identifier of group

Please provide your assessment of each report in the following fields according to the scales presented above:

Threats Quality	Threats quality inline with the corresponding scale
Security Controls Quality	Security Controls quality inline with the corresponding scale
Comments	Here you can provide participants and us your comments to the report

Table 7.4: Post-task Questionnaire

Table reports post-task questions and their perception type, PU or PEOU (questions about intention to use and perceive leverage are omitted). Some questions do no specify whether the method was used for threats or for controls. In that case we have used the corresponding answers for both threats and controls.

Q#	**Type**	**Question (positive statement)**
1	PEOU	SecRAM helped me in brainstorming on the threats
2	PEOU	SecRAM helped me in brainstorming on the security controls
3	PEOU	I found SecRAM easy to use
4	PU	SecRAM process is well detailed
5	PEOU	SecRAM was difficult to master
6	PEOU	I was never confused about how to apply SecRAM to the application
7	PU	I would have found specific threats more quickly with the SecRAM
8	PU	I would have found specific security controls more quickly with the SecRAM
9	PU	SecRAM made the security analysis more systematic
10	PEOU	SecRAM made it easier to evaluate whether threats were appropriate to the context
11	PEOU	SecRAM made it easier to evaluate whether security controls were appropriate to the context
12	PU	SecRAM made the search for specific threats more systematic
13	PU	SecRAM made the search for specific security controls more systematic
14	PU	If I need to update the analysis it will be easier with SecRAM than with common sense
15	PU	SecRAM made the security analysis easier than an ad hoc approach
16	PU	SecRAM made me more productive in finding threats
17	PU	SecRAM made me more productive in finding security controls

Table 7.5: Participants, Their Results and Quality Assessment – Experiment 3.1 (Novices (Students))

Table presents the information about number of threats and security controls identified by participants and the assessment from three ATM experts on the quality of threats and security controls. Note: T – threats, SC – security controls.

ID	Catalog	Quantity		Quality					
				Expert 1		Expert 2		Expert 3	
		T	SC	T	SC	T	SC	T	SC
G01	DOM CAT	17	32	3	3	2	2	2	2
G02	DOM CAT	53	61	4	3	3	3	3	3
G03	DOM CAT	35	145	4	4	4	4	4	4
G04	DOM CAT	28	55	4	3	3	4	3	3
G05	DOM CAT	15	16	3	3	3	3	5	5
G06	GEN CAT	18	42	3	4	3	3	4	4
G07	GEN CAT	36	26	3	4	3	4	3	3
G08	GEN CAT	44	44	2	3	4	4	3	3
G09	GEN CAT	30	33	3	4	3	3	3	4

Table 7.6: Participants, Their Results and Quality Assessment – Experiment 3.2 (ATM Professionals)

Table presents the information about security knowledge, working experience and degree of participants; number of threats and security controls identified by participants and the assessment from two ATM experts on the quality of threats and security controls. Note: T – threats, SC – security controls.

ID	Security Knowl.	Working Exp.	Education Degree	Catalog	Quantity		Quality			
							Expert 1		Expert 2	
					T	SC	T	SC	T	SC
P01	No	6	MSC	GEN CAT	17	28	2	2	3	3
P02	No	5	PHD	GEN CAT	9	17	1	2	2	2
P03	Yes	4	MSC	GEN CAT	27	50	4	4	4	3
P04	No	5	MSC	GEN CAT	9	23	2	2	3	3
P05	Yes	4	PHD	GEN CAT	9	15	3	3	3	3
P06	No	8	DIPLOMA	DOM CAT	22	38	4	3	3	3
P07	No	4	MSC	DOM CAT	7	14	2	2	2	2
P08	No	5	PHD	DOM CAT	24	66	4	4	4	4
P09	Yes	2	MSC	DOM CAT	24	45	5	4	5	4
P10	No	7	PHD	DOM CAT	16	32	4	4	3	3
P11	No	5	MSC	NO CAT	10	13	2	1	3	3
P12	Yes	14	PHD	NO CAT	15	47	3	3	4	3
P13	Yes	17	MSC	NO CAT	15	19	2	3	3	3
P14	Yes	18	MSC	NO CAT	24	28	2	2	3	3
P15	Yes	15	MSC	NO CAT	6	13	2	4	4	3

Table 7.7: Responses to the Post-task Questions – Experiment 3.1

Table reports mean and median value of participants' responses in the first experiment to each post-task question and the type of the question.

Q#	Type	DOM CAT		GEN CAT	
		Mean	Median	Mean	Median
1	PEOU	3.4	4	3.625	3.5
2	PEOU	3.7	4	3.5	3.5
3	PEOU	3.5	4	4.25	4.5
4	PU	3.2	3.5	2.625	2.5
5	PEOU	3.8	4	3.875	4
6	PEOU	3.1	3.5	3.375	3.5
7	PU	3.5	3	3	3
8	PU	3.8	4	3.25	3
9	PU	3.9	4	3.625	4
10	PEOU	3.7	4	2.75	3
11	PEOU	3.7	4	3	3
12	PU	3.8	4	3.25	3
13	PU	3.8	4	3.625	4
14	PU	3.8	4	2.875	3
15	PU	3.6	3	3.125	3
16	PU	3.7	4	3	3
17	PU	3.9	4	3.375	3

Table 7.8: Responses to the Post-task Questions – Experiment 3.2

Table reports mean and median value of participants' responses in the second experiment to each post-task question and the type of the question.

Q#	Type	DOM CAT		GEN CAT		NO CAT	
		Mean	Median	Mean	Median	Mean	Median
1	PEOU	4.2	4	4	4	3.2	3
2	PEOU	4.2	4	3.2	4	3.2	3
3	PEOU	3.4	3	3.2	4	4.2	4
4	PU	3.4	4	3.4	3	3.8	4
5	PEOU	3	3	3.4	4	3.8	4
6	PEOU	2.8	3	2.6	3	4	4
7	PU	3.4	3	2.4	2	3.2	3
8	PU	3.8	4	2.4	2	3.2	3
9	PU	3.8	4	4.2	4	4.2	5
10	PEOU	3.2	3	3.4	4	3	3
11	PEOU	2.8	3	2.6	2	3	3
12	PU	3.8	4	3.8	4	3.6	3
13	PU	3.4	3	3.6	4	3.6	4
14	PU	4	4	3.6	4	4.6	5
15	PU	2.8	3	2.6	3	3.6	4
16	PU	4.2	4	3	4	3.4	4
17	PU	4	4	3.4	4	3.4	3

Chapter 8

Additional Data for Chapter 4

Table 8.1: Post-Task Questionnaire

This is the post-task questionnaire that we distributed to the participants. Questions $Q1$-$Q8$ included closed answers on a 5-point Likert scale: 0 – strongly agree, 1 – agree, 2 – not certain, 3 – disagree, and 4 – strongly disagree. Only question $Q9$ had "yes" and "no" answers.

Q#	Statement
Q1	I had enough time to perform the task
Q2	The objectives of the study were perfectly clear to me
Q3	The task I had to perform was perfectly clear to me
Q4	The comprehensibility questions were perfectly clear to me
Q5	I experienced no difficulty to answer the comprehensibility questions
Q6	I experienced no difficulty in understanding the risk model tables (diagrams)
Q7	I experienced no difficulty in using electronic version of the risk model tables (diagrams)
Q8	I experienced no difficulty in using SurveyGizmo
Q9	[Tabular] Did you use search, or filtering, or sorting function in Excel or OpenOffice document? [Graphical] Did you use search in the PDF document?

Threat Event	Threat Source	Vulnerabilities	Impact	Asset	Overall Likelihood	Level of Impact	Security Controls
Error in the role assignment leads to elevation of privilege.	Admin	Insufficient routines	Unauthorized data modification	Data integrity	Unlikely	Severe	1. Strengthen routines for access control policy specification. 2. Conduct regular audits of assigned user roles.
Error in the role assignment leads to elevation of privilege.	Admin	Insufficient routines	Unauthorized data access	Data confidentiality	Likely	Severe	1. Strengthen routines for access control policy specification. 2. Conduct regular audits of assigned user roles.
Error in the role assignment leads to elevation of privilege.	Admin	Insufficient routines	Unauthorized data access	Privacy	Likely	Critical	1. Strengthen routines for access control policy specification. 2. Conduct regular audits of assigned user roles.
SQL injection attack leads to successful SQL injection.	Hacker	Insufficient input validation	Unauthorized data access	Data confidentiality	Likely	Severe	Implement strong input validation.
SQL injection attack leads to successful SQL injection.	Hacker	Insufficient input validation	Unauthorized data access	Privacy	Likely	Critical	Implement strong input validation.
SQL injection attack leads to successful SQL injection.	Hacker	Insufficient input validation	Unauthorized data modification	Data integrity	Unlikely	Severe	Implement strong input validation.
Error in assignment of privacy level leads to insufficient data anonymization.	Data reviewer	Insufficient routines	Unauthorized access to personal identifiable information	Privacy	Unlikely	Critical	Strengthen routines for privacy level specification.
Cyber criminal sends crafted phishing emails to HCN users and this leads to sniffing of user credentials.	Cyber criminal	1. Lack of security awareness 2. Weak authentication	Unauthorized access to HCN	Data confidentiality	Very likely	Critical	1. Improve security training. 2. Strengthen authentication mechanism.
Cyber criminal sends crafted phishing emails to HCN users and this leads to sniffing of user credentials.	Cyber criminal	1. Lack of security awareness 2. Weak authentication	Unauthorized access to HCN	Privacy	Very likely	Critical	1. Improve security training. 2. Strengthen authentication mechanism.
Cyber criminal sends crafted phishing emails to HCN users and this leads to that HCN network infected by malware.	Cyber criminal	Lack of security awareness	Leakage of patient data	Privacy	Very unlikely	Critical	Improve security training.
Cyber criminal sends crafted phishing emails to HCN users and this leads to that HCN network infected by malware.	Cyber criminal	Lack of security awareness	Leakage of patient data	Data confidentiality	Very unlikely	Severe	Improve security training.
HCN user connects private mobile device to the network and this leads to that HCN network infected by malware.	HCN user	1. Insufficient security policy 2. Insufficient malware detection	Leakage of patient data	Privacy	Very unlikely	Critical	1. Impose security policy on the use of mobile devices. 2. Implement state-of-the-art malware detection.
HCN user connects private mobile device to the network and this leads to that HCN network infected by malware.	HCN user	1. Insufficient security policy 2. Insufficient malware detection	Leakage of patient data	Data confidentiality	Very unlikely	Severe	1. Impose security policy on the use of mobile devices. 2. Implement state-of-the-art malware detection.

Figure 8.1: Risk Model for HCN Scenario in Tabular Notation Provided to the Participants

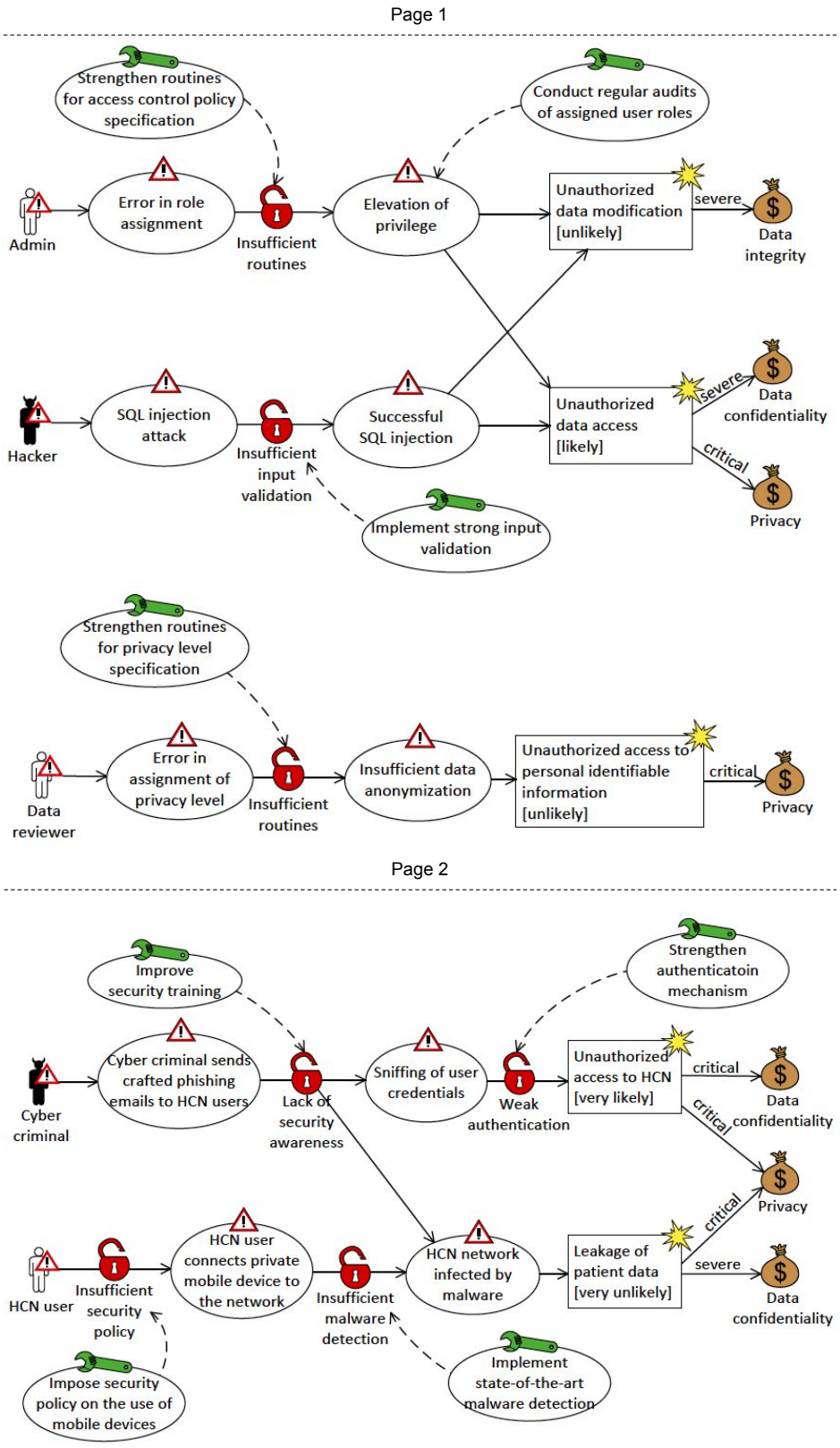


Figure 8.2: Risk Model for HCN Scenario in Graphical Notation Provided to the Participants

Table 8.2: Comprehension Questions for Graphical Risk Model – Study 4.1

This table presents the exact comprehension questionnaire that we provided to the participants of the first study with graphical risk model.

Q#	Complexity	Question statement
1	2	Which threat scenarios can be initiated by exploiting vulnerability "Insufficient routines", according to the risk model? Please list all threat scenarios:
2	4	Which unwanted incidents are possible as a result of exploiting vulnerability "Lack of security awareness" by Cyber criminal? Specify all unwanted incidents:
3	2	Which are the assets that can be harmed by the unwanted incident "Unauthorized access to HCN"? Please list all assets:
4	2	What is the likelihood that unwanted incident "Unauthorized data access" occurs? Specify the likelihood:
5	6	What is the highest possible consequence for the asset "Data confidentiality" that Cyber criminal or Hacker can cause? Please specify the consequence:
6	2	Which threats can exploit the vulnerability "Insufficient routines"? Please specify all threats:
7	3	What are the vulnerabilities that can be exploited to initiate each of the following threat scenarios: "HCN network infected by malware" and "Elevation of privilege"? Please list all vulnerabilities:
8	4	Which treatments are used to mitigate vulnerabilities "Insufficient routines" or threat scenario "Elevation of privilege"? Please specify all treatments:
9	2	Which threats can attack the asset "Privacy"? Please specify all threats:
10	4	Which threat scenarios can Cyber criminal initiate to harm the asset "Data confidentiality"? Please list all threat scenarios:
11	4	Which treatments can be used to mitigate vulnerabilities exploited by Cyber criminal to attack the asset "Privacy"? Please list all treatments:
12	6	Which are the unwanted incidents that can be initiated by Hacker or Cyber criminal and can occur, according to the risk model? Please list all unwanted incidents:

Table 8.3: Comprehension Questions for Graphical Risk Model – Study 4.2

This table presents the exact comprehension questionnaire that we provided to the participants of the second study with graphical risk model.

$Q\#$	IC	R	J	Question statement
1	1	1	-	What are the consequences that can be caused for the asset "Availability of service"? Please specify the consequences that meet the conditions.
2	1	1	-	Which vulnerabilities can lead to the unwanted incident "Unauthorized transaction via Poste App"? Please list all vulnerabilities that meet the conditions.
3	2	1	-	Which assets can be impacted by Hacker or System failure? Please list all unique assets that meet the conditions.
4	2	1	-	Which unwanted incidents can be initiated by Cyber criminal with consequence equal to "sever"? Please list all unwanted incidents that meet the conditions.
5	2	2	-	Which threat scenarios can be initiated by Cyber criminal to impact the asset "Confidentiality of customer data"? Please list all unique threat scenarios that meet the conditions.
6	2	2	-	Which treatments can be used to mitigate attack paths caused by any of the vulnerabilities "Poor security awareness" or "Lack of mechanisms for authentication of app"? Please list all unique treatments for all attack paths caused by any of the specified vulnerabilities.
7	1	1	1	What is the lowest consequence that can be caused for the asset "User authenticity"? Please specify the consequence that meet the conditions.
8	1	1	1	Which threats can impact assets with consequence equal to "severe" or higher? Please list all threats that meet the conditions.
9	2	1	1	Which unwanted incidents can be initiated by Hacker with likelihood equal to "likely" or higher? Please list all unwanted incidents that meet the conditions.
10	2	1	1	What is the lowest likelihood of the unwanted incidents that can be caused by any of the vulnerabilities "Use of web application" or "Poor security awareness"? Please specify the lowest likelihood of the unwanted incidents that can be initiated using any of the specified vulnerabilities.
11	2	2	1	Which vulnerabilities can be exploited by Hacker to initiate unwanted incidents with likelihood equal to "likely" or higher? Please list all vulnerabilities that meet the conditions.
12	2	2	1	What is the lowest consequence of the unwanted incidents that can be caused by Hacker and mitigated by treatment "Regularly inform customers of security best practices"? Please specify the lowest consequence that meets the conditions.

Table 8.4: Precision and Recall by Questions – Study 4.1

The most significant difference (≥ 0.2) in precision was observed for Q1, Q6 and in recall for Q2, Q5-Q7, and Q10. In all these questions tabular models showed better results. Column "∅" reports the number of empty responses to a question which can be caused by task termination forced by SurveyGizmo due to time limit.

Q#	Comp-	Tabular					Graphical				
	lexity	#obs.	∅	mean	med.	sd	#obs.	∅	mean	med.	sd
Precision											
Q1	2	33	0	1.00	1.00	0.00	36	0	0.79	1.00	0.37
Q2	4	33	0	0.92	1.00	0.25	36	0	0.81	1.00	0.40
Q3	2	33	0	0.99	1.00	0.06	36	0	0.95	1.00	0.19
Q4	2	33	0	0.94	1.00	0.24	36	0	0.86	1.00	0.35
Q5	6	33	0	0.64	1.00	0.46	36	0	0.46	0.25	0.48
Q6	2	33	0	0.99	1.00	0.06	36	0	0.66	1.00	0.44
Q7	4	33	0	0.97	1.00	0.10	36	0	0.94	1.00	0.20
Q8	4	33	0	0.99	1.00	0.06	36	0	0.96	1.00	0.18
Q9	2	33	0	0.94	1.00	0.24	36	0	0.88	1.00	0.32
Q10	4	33	0	0.87	1.00	0.27	36	0	0.85	1.00	0.31
Q11	4	33	0	0.83	1.00	0.29	36	0	0.85	1.00	0.31
Q12	6	33	0	0.53	0.50	0.27	36	0	0.61	0.50	0.35
Overall		33	0	0.88	1.00	0.27	36	0	0.80	1.00	0.36
Recall											
Q1	2	33	0	0.97	1.00	0.12	36	0	0.79	1.00	0.37
Q2	4	33	0	0.92	1.00	0.25	36	0	0.61	0.5	0.38
Q3	2	33	0	1.00	1.00	0.00	36	0	0.96	1.00	0.18
Q4	2	33	0	0.94	1.00	0.24	36	0	0.86	1.00	0.35
Q5	6	33	0	0.70	1.00	0.47	36	0	0.50	0.5	0.51
Q6	2	33	0	0.95	1.00	0.15	36	0	0.65	1.00	0.44
Q7	4	33	0	0.89	1.00	0.20	36	0	0.62	0.75	0.24
Q8	4	33	0	0.80	0.67	0.17	36	0	0.78	1.00	0.28
Q9	2	33	0	0.87	1.00	0.26	36	0	0.73	0.80	0.32
Q10	4	33	0	0.91	1.00	0.23	36	0	0.66	0.67	0.30
Q11	4	33	0	0.98	1.00	0.09	36	0	0.89	1.00	0.27
Q12	6	33	0	0.80	1.00	0.35	36	0	0.79	1.00	0.38
Overall		33	0	0.90	1.00	0.25	36	0	0.74	1.00	0.36

Table 8.5: Precision and Recall by Questions – Study 4.2

The most significant difference (≥ 0.2) in precision was revealed for Q1, Q8, Q10, and Q12, and in recall of almost half of the questions (Q1, Q4-Q6,Q8,Q10, and Q12). For all these questions tabular model showed better results than the graphical one. Column "∅" reports the number of empty responses to a question which can be caused by task termination forced by SurveyGizmo due to time limit.

Q#	Comp-	Tabular					Graphical				
	lexity	#obs.	∅	mean	med.	sd	#obs.	∅	mean	med.	sd
Precision											
Q1	2	83	1	0.95	1.00	0.22	83	0	0.64	1.00	0.48
Q2	2	83	1	0.95	1.00	0.22	83	1	0.95	1.00	0.20
Q3	3	83	1	1.00	1.00	0.04	83	0	0.99	1.00	0.07
Q4	3	83	0	0.95	1.00	0.20	83	2	0.90	1.00	0.29
Q5	4	83	0	0.99	1.00	0.07	83	0	0.90	1.00	0.28
Q6	4	83	0	1.00	1.00	0.03	83	0	0.99	1.00	0.08
Q7	3	83	2	0.89	1.00	0.32	83	0	0.73	1.00	0.44
Q8	3	83	1	0.97	1.00	0.15	83	0	0.71	1.00	0.44
Q9	4	83	1	0.85	1.00	0.29	83	0	0.88	1.00	0.24
Q10	4	83	1	0.65	1.00	0.48	83	1	0.43	0.00	0.50
Q11	5	83	0	0.93	1.00	0.19	83	0	0.84	1.00	0.32
Q12	5	83	1	0.85	1.00	0.36	83	0	0.64	1.00	0.48
Overall		83	9	0.91	1.00	0.26	83	4	0.80	1.00	0.39
Recall											
Q1	2	83	1	0.95	1.00	0.22	83	0	0.64	1.00	0.48
Q2	2	83	1	0.94	1.00	0.23	83	1	0.76	1.00	0.28
Q3	3	83	1	1.00	1.00	0.00	83	0	0.96	1.00	0.14
Q4	3	83	0	0.87	1.00	0.25	83	2	0.63	0.67	0.29
Q5	4	83	0	0.94	1.00	0.15	83	0	0.64	0.75	0.32
Q6	4	83	0	0.86	1.00	0.17	83	0	0.60	0.60	0.20
Q7	3	83	2	0.89	1.00	0.32	83	0	0.73	1.00	0.44
Q8	3	83	1	0.97	1.00	0.14	83	0	0.64	0.67	0.42
Q9	4	83	1	0.77	1.00	0.32	83	0	0.81	1.00	0.29
Q10	4	83	1	0.65	1.00	0.48	83	1	0.43	0.00	0.50
Q11	5	83	0	0.84	1.00	0.25	83	0	0.67	0.50	0.32
Q12	5	83	1	0.85	1.00	0.36	83	0	0.64	1.00	0.48
Overall		83	9	0.88	1.00	0.28	83	4	0.68	1.00	0.38

8.1 Effect of Task Complexity Components on the Risk Model Comprehension

Figure 8.3 shows the interaction plots between F-measure by model type (graphical vs. tabular) and the levels of IC.

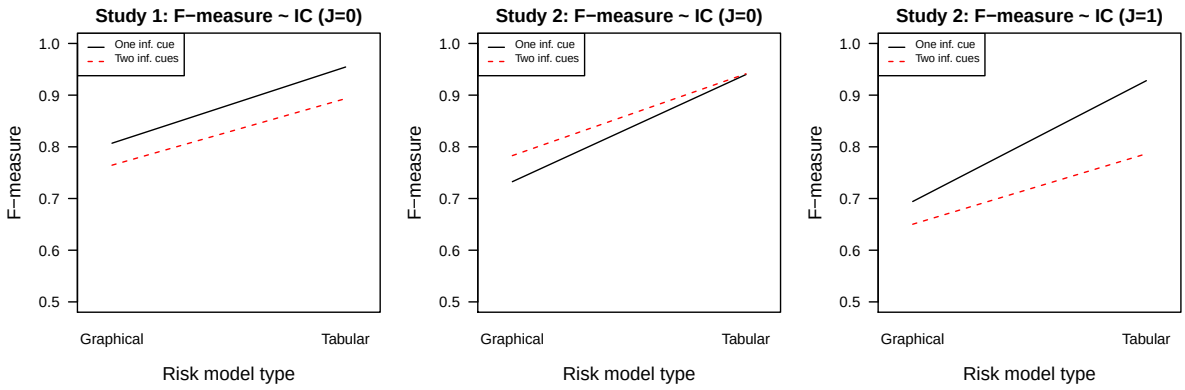


Figure 8.3: Effect of Complexity (IC) on F-measure

Figure 8.4 shows the interaction plots between F-measure by model type (graphical vs. tabular) and the levels of R.

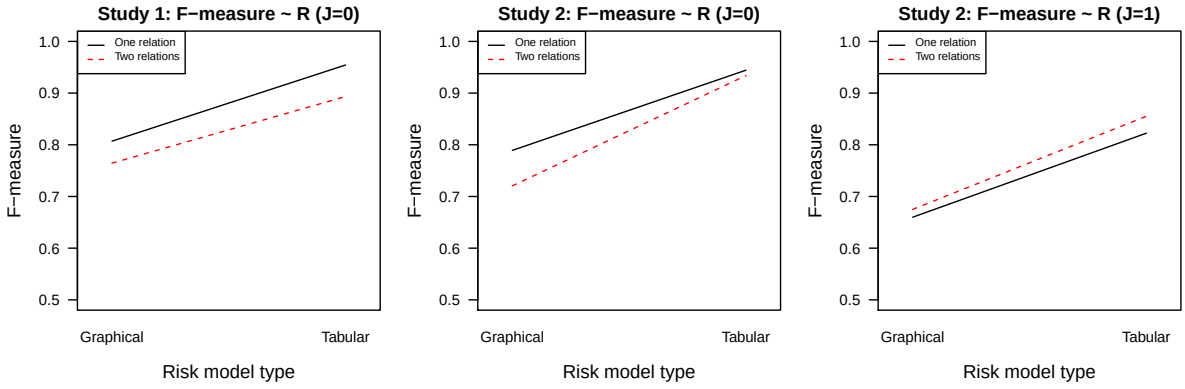


Figure 8.4: Effect of Complexity (R) on F-measure

Figure 8.5 shows the interaction plots between F-measure by model type (graphical vs. tabular) and the presence of the judgment component.

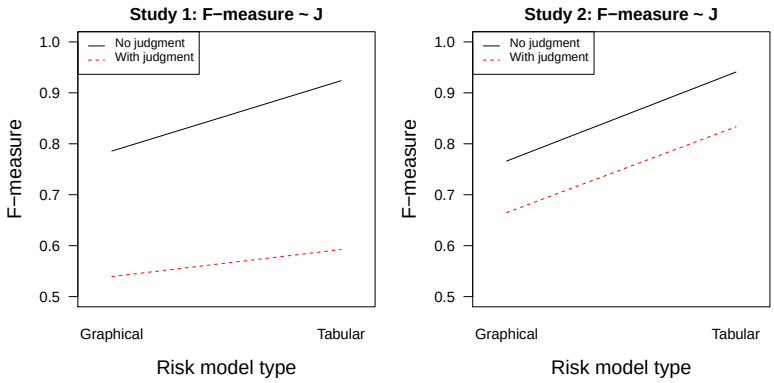


Figure 8.5: Effect of Complexity (J) on F-measure

Bibliography

[1] Silvia Abrahao, Carmine Gravino, Emilio Insfran, Giuseppe Scanniello, and Genoveffa Tortora. Assessing the effectiveness of sequence diagrams in the comprehension of functional requirements: Results from a family of five experiments. *IEEE Transactions on Software Engineering*, 39(3):327–342, 2013.

[2] Ritu Agarwal, Prabuddha De, and Atish P. Sinha. Comprehending object and process models: An empirical study. *IEEE Transactions on Software Engineering*, 25(4):541–556, 1999.

[3] Sean Barnum and Gary McGraw. Knowledge for software security. *IEEE Security & Privacy*, 3(2):74–78, 2005.

[4] Victor R Basili and H Dieter Rombach. The TAME project: towards improvement-oriented softwareenvironments. *IEEE Transactions on Software Engineering*, 14(6):758–773, 1988.

[5] BSI. Standard 100-1: Information security management systems. 2012.

[6] Mark Chatfield and Adrian Mander. The skillings–mack test (friedman test when there are missing data). *Stata Journal*, 9(2):299, 2009.

[7] Lawrence Chung, Brian A. Nixon, Eric Yu, and John Mylopoulos. *Non-functional requirements in Software Engineering*. Kluwer Academic Publishers, 2000.

[8] Nelly Condori-Fernandez, Maya Daneva, Klaas Sikkel, Roel Wieringa, Oscar Dieste, and Oscar Pastor. A systematic mapping study on empirical evaluation of software requirements specifications techniques. In *Proceeding of the 3rd International Symposium on Empirical Software Engineering and Measurement*, pages 502–505. IEEE, 2009.

[9] William Jay Conover. On methods of handling ties in the wilcoxon signed-rank test. *Journal of the American Statistical Association*, 68(344):985–988, 1973.

[10] Francisco Lopez Crespo, Miguel Angel Amutio Gomez, Javier Candau, and Jose Antonio Manas Manas. Magerit: Methodology for information systems risk analysis and management. Technical report, Ministerio de Administraciones Publicas, 2006.

[11] Zapata Belén Cruz, José Luis Fernández-Alemán, and Ambrosio Toval. Security in cloud computing: a mapping study. *Computer Science and Information Systems*, 12(1):161–184, 2015.

[12] Luiz Marcio Cysneiros. Evaluating the effectiveness of using catalogues to elicit non-functional requirements. In *Proceedings of the 10th Workshop on Requirements Engineering*, pages 107–115, 2007.

[13] Olawande Daramola, Yushan Pan, Péter Kárpáti, and Guttorm Sindre. A comparative review of i*-based and use case-based security modelling initiatives. In *Proceeding of the 6th International Conference on Research Challenges in Information Science*, pages 1–12. IEEE, 2012.

[14] Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information Systems Quarterly*, pages 319–340, 1989.

[15] Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. Extrinsic and intrinsic motivation to use computers in the workplace1. *Journal of Applied Social Psychology*, 22(14):1111–1132, 1992.

[16] Martina de Gramatica, Katsiaryna Labunets, Fabio Massacci, Federica Paci, and Alessandra Tedeschi. The Role of Catalogues of Threats and Security Controls in Security Risk Assessment: An Empirical Study with ATM Professionals. In *Proceedings of the 21st International Working Conference on Requirements Engineering: Foundation for Software Quality*, volume 9013 of *Lecture Notes in Computer Science*, pages 98–114. Springer, 2015.

[17] Andrea De Lucia, Carmine Gravino, Rocco Oliveto, and Genoveffa Tortora. An experimental comparison of ER and UML class diagrams for data modelling. *Empirical Software Engineering*, 15(5):455–492, 2010.

[18] Mina Deng, Kim Wuyts, Riccardo Scandariato, Bart Preneel, and Wouter Joosen. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering*, 16(1):3–32, 2011.

[19] David Dunning, Kerri Johnson, Joyce Ehrlinger, and Justin Kruger. Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 12(3):83–87, 2003.

[20] Sergio España, Nelly Condori-Fernandez, Arturo González, and Óscar Pastor. An empirical comparative evaluation of requirements engineering methods. *Journal of the Brazilian Computer Society*, 16(1):3–19, 2010.

[21] EUROCONTROL. *EATM, ATM Security Risk Assessment Methodology, Edition 1.0*, May 2008.

[22] Benjamin Fabian, Seda Gürses, Maritta Heisel, Thomas Santen, and Holger Schmidt. A comparison of security requirements engineering methods. 15(1):7–40, 2010.

[23] Davide Falessi, Lionel C Briand, Giovanni Cantone, Rafael Capilla, and Philippe Kruchten. The value of design rationale information. 22(3):21, 2013.

[24] Thomas Falk, Philipp Griesberger, Florian Johannsen, and Susanne Leist. Patterns for Business Process Improvement – A First Approach. In *Proceedings of the 21st European Conference on Information Systems*, 2013.

[25] Samer Faraj and Lee Sproull. Coordinating expertise in software development teams. *Management Science*, 46(12):1554–1568, 2000.

[26] Food and Drug Administration. Guidance for industry: Statistical approaches to establishing bioequivalence, 2001.

[27] Gordon Fraser and Andrea Arcuri. Sound empirical evidence in software testing. In *Proceedings of the 34th International Conference on Software Engineering*, pages 178–188. IEEE Press, 2012.

[28] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. *Design patterns: elements of reusable object-oriented software*. Pearson Education, 1994.

[29] Shang Gao, John Krogstie, and Keng Siau. Adoption of mobile information services: An empirical study. *Mobile Information Systems*, 10(2):147–171, 2014.

[30] Luis Garicano and Yanhui Wu. Knowledge, communication, and organizational capabilities. *Organization Science*, 23(5):1382–1397, 2012.

[31] Paolo Giorgini, Fabio Massacci, John Mylopoulos, and Nicola Zannone. Modeling security requirements through ownership, permission and delegation. In *Proceedings of the 13th IEEE International Conference on Requirements Engineering*, pages 167–176. IEEE, 2005.

[32] Barney G. Glaser and Anselm L. Strauss. *The Discovery of Grounded Theory*. Transaction Publishers, 1967.

[33] Milos Gligoric, Alex Groce, Chaoqiang Zhang, Rohan Sharma, Mohammad Amin Alipour, and Darko Marinov. Comparing non-adequate test suites using coverage criteria. In *Proceedings of the 22nd International Symposium on Software Testing and Analysis*, pages 302–313. ACM, 2013.

[34] Shirley Gregor. The nature of theory in information systems. *Management Information Systems Quarterly*, pages 611–642, 2006.

[35] Ida Hogganvik Grøndahl, Mass Soldal Lund, and Ketil Stølen. Reducing the effort to comprehend risk models: Text labels are often preferred over graphical means. *Risk Analysis*, 31:1813–1831, 2011.

[36] Greg Guest, Kathleen M. MacQueen, and Emily E. Namey. *Applied thematic analysis*. Sage, 2011.

[37] Irit Hadar, Iris Reinhartz-Berger, Tsvi Kuflik, Anna Perini, Filippo Ricca, and Angelo Susi. Comparing the comprehensibility of requirements models expressed in use case and tropos: Results from a family of experiments. *Information and Software Technology*, 55(10):1823–1843, 2013.

[38] Mohanad Halaweh. Using grounded theory as a method for system requirements analysis. *Journal of Information Systems and Technology Management*, 9(1):23–38, 2012.

[39] C.B. Haley, R. Laney, J.D. Moffett, and B. Nuseibeh. Security requirements engineering: A framework for representation and analysis. *IEEE Transactions on Software Engineering*, 34:133–153, 2008.

[40] Motulsky Harvey. Intuitive biostatistics, 1995.

[41] Werner Heijstek, Thomas Kühne, and Michel RV Chaudron. Experimental analysis of textual and graphical representations for software architecture design. In *Proceedings of the 5th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 167–176. IEEE, 2011.

[42] Robert R. Hoffman, Nigel R. Shadbolt, A. Mike Burton, and Gary Klein. Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes*, 62(2):129–158, 1995.

[43] Ida Hogganvik and Ketil Stølen. On the comprehension of security risk scenarios. pages 115–124. IEEE, 2005.

[44] Bernhard Hoisl, Stefan Sobernig, and Mark Strembeck. Comparing three notations for defining scenario-based model tests: A controlled experiment. In *Proceedings of the 9th International Conference on the Quality of Information and Communications Technology*, pages 95–104. IEEE, 2014.

[45] Michael Jackson. *Problem frames: analysing and structuring software development problems*. Addison-Wesley, 2001.

[46] Natalia Juristo and Ana M Moreno. *Basics of software engineering experimentation*. Springer, 2010.

[47] Robert Kabacoff. *R in action: data analysis and graphics with R*. Manning Publications Co., 2015.

[48] Monika Kaczmarek, Alexander Bock, and Michael Heß. On the explanatory capabilities of enterprise modeling approaches. In *Proceedings of the 5th Enterprise Engineering Working Conference*, pages 128–143. Springer, 2015.

[49] Elena Karahanna, Detmar W Straub, and Norman L Chervany. Information technology adoption across time: a cross-sectional comparison of pre-adoption and post-adoption beliefs. *Management Information Systems Quarterly*, pages 183–213, 1999.

[50] Peter Karpati, Andreas L. Opdahl, and Guttorm Sindre. Investigating security threats in architectural context: Experimental evaluations of misuse case maps. *Journal of Systems and Software*, 104:90–111, 2015.

[51] Peter Karpati, Yonathan Redda, Andreas L. Opdahl, and Guttorm Sindre. Comparing attack trees and misuse cases in an industrial setting. *Information and Software Technology*, 56(3):294–308, 2014.

[52] Peter Karpati, Guttorm Sindre, and Andreas L Opdahl. Characterising and analysing security requirements modelling initiatives. In *Proceedings of the 6th IEEE International Conference on Availability, Reliability and Security*, pages 710–715. IEEE, 2011.

[53] Azmeri Khan and Glen D. Rayner. Robustness to non-normality of common tests for the many-sample location problem. *Journal of Applied Mathematics and Decision Sciences*, 7(4):187–206, 2003.

[54] Humera Khan and PDD Dominic. User acceptance of online system: a study of banking and airline sector. *International Journal of Business Information Systems*, 16(4):359–374, 2014.

[55] Katsiaryna Labunets, Fabio Massacci, Federica Paci, and Le Minh Sang Tran. An Experimental Comparison of Two Risk-Based Security Methods. In *Proceedings of the 7th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 163–172. IEEE, 2013.

[56] Katsiaryna Labunets, Federica Paci, and Fabio Massacci. Which Security Catalogue Is Better for Novices? In *Proceedings of the 5th IEEE International Workshop on Empirical Requirements Engineering at the 23rd IEEE International Requirements Engineering Conference*, pages 25–32, 2015.

[57] Katsiaryna Labunets, Federica Paci, Fabio Massacci, Martina Ragosta, and Bjørnar Solhaug. A First Empirical Evaluation Framework for Security Risk Assessment Methods in the ATM Domain. In *Proceedings of the 4th SESAR Innovation Days*. SESAR, 2014.

[58] Katsiaryna Labunets, Federica Paci, Fabio Massacci, and Raminder Ruprai. An experiment on comparing textual vs. visual industrial methods for security risk assessment. In *Proceedings of the 4th IEEE International Workshop on Empirical Requirements Engineering at the 22nd IEEE International Requirements Engineering Conference*, pages 28–35. IEEE, 2014.

[59] Douglas J Landoll and Douglas Landoll. *The security risk assessment handbook: A complete guide for performing security risk assessments*. CRC Press, 2005.

[60] Øyvind Langsrud. Anova for unbalanced data: Use type ii instead of type iii sums of squares. *Statistics and Computing*, 13(2):163–167, 2003.

[61] Mass Soldal Lund, Bjørnar Solhaug, and Ketil Stølen. A guided tour of the CORAS method. In *Model-Driven Risk Analysis*, pages 23–43. Springer, 2011.

[62] I. Scott MacKenzie. *Human-computer interaction: An empirical research perspective*. Newnes, 2012.

[63] Lynne M Markus. Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems*, 18(1):57–93, 2001.

[64] Yulkeidi Martínez, Cristina Cachero, and Santiago Meliá. Mdd vs. traditional software development: A practitioner's subjective perspective. *Information and Software Technology*, 2012.

[65] Fabio Massacci and Federica Paci. How to select a security requirements method? a comparative study with students and practitioners. In *Proceedings of the 17th Nordic Conference on Secure IT Systems*, pages 89–104. Springer, 2012.

[66] Raimundas Matulevičius, Nicolas Mayer, Haralambos Mouratidis, Eric Dubois, Patrick Heymans, and Nicolas Genon. Adapting secure tropos for security risk management in the early phases of information systems development. pages 541–555. Springer, 2008.

[67] Alistair Mavin and Neil Maiden. Determining socio-technical systems requirements: experiences with generating and walking through scenarios. In *Proceedings of the 11th IEEE International Conference on Requirements Engineering*, pages 213–222. IEEE, 2003.

[68] Nicolas Mayer, Patrick Heymans, and Raimundas Matulevicius. Design of a modelling language for information system security risk management. pages 121–132, 2007.

[69] Nicolas Mayer, André Rifaut, and Eric Dubois. Towards a risk-based security requirements engineering framework. volume 5, 2005.

[70] Nancy R Mead, Julia H Allen, Sean Barnum, Robert J Ellison, and Gary McGraw. *Software Security Engineering: A Guide for Project Managers*. Addison-Wesley Professional, 2004.

[71] Daniel Mellado, Eduardo Fernández-Medina, and Mario Piattini. Applying a security requirements engineering process. In *Proceeding of the 11th European Symposium on Research in Computer Security*, pages 192–206. Springer, 2006.

[72] Daniel Mellado, Eduardo Fernández-Medina, and Mario Piattini. Towards security requirements management for software product lines: A security domain requirements engineering process. *Computer Standards and Interfaces*, 30(6):361–371, 2008.

[73] Daniel Mellado and David G Rosado. An overview of current information systems security challenges and innovations. *Journal of Universal Computer Science*, 18(12):1598–1607, 2012.

[74] J. Patrick Meyer and Michael A Seaman. A comparison of the exact Kruskal-Wallis distribution to asymptotic approximations for all sample sizes up to 105. *Journal of Experimental Education*, 81(2):139–156, 2013.

[75] Michael Meyners. Equivalence tests–a review. *Food quality and preference*, 26(2):231–245, 2012.

[76] Daniel Moody. The "Physics" of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. *IEEE Transactions on Software Engineering*, 35(6):756–779, 2009.

[77] Daniel L. Moody. The method evaluation model: A theoretical model for validating information systems design methods. In *Proceedings of the 11th European Conference of Information Systems*, pages 1327–1336, 2003.

[78] Mirko Morandini, Anna Perini, and Alessandro Marchetto. Empirical evaluation of Tropos4AS modelling. In *Proceedings of the 5th International i* Workshop*, volume 766, pages 14–19. CEUR, 2011.

[79] Adrien Mouaffo, Davide Taibi, and Kavyashree Jamboti. Controlled experiments comparing fault-tree-based safety analysis techniques. In *Proceedings of the 18th*

International Conference on Evaluation and Assessment in Software Engineering, page 46. ACM, 2014.

[80] Samar Mouakket. The motivations for citizens' adoption of e-government: an empirical study in the uae. *International Journal of Business Information Systems*, 6(2):240–264, 2010.

[81] Haralambos Mouratidis and Paolo Giorgini. Secure tropos: a security-oriented extension of the tropos methodology. *International Journal of Software Engineering and Knowledge Engineering*, 17(02):285–309, 2007.

[82] Haralambos Mouratidis, Paolo Giorgini, and Gordon Manson. Integrating security and systems engineering: Towards the modelling of secure information systems. In *Proceedings of the 15th International Conference on Advanced Information Systems Engineering*, volume 2681, pages 1031–1031. Springer, 2003.

[83] Elias Mturi and Paul Johannesson. A context-based process semantic annotation model for a process model repository. *Business Process Management Journal*, 19(3):404–430, 2013.

[84] Nadim Nachar. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4(1):13–20, 2008.

[85] Armstrong Nhlabatsi, Bashar Nuseibeh, and Yijun Yu. Security requirements engineering for evolving software systems: A survey. *International Journal of Secure Software Engineering*, 1(1):54–73, 2010.

[86] Andreas L Opdahl and Guttorm Sindre. Experimental comparison of attack trees and misuse cases for security threat identification. *Information and Software Technology*, 51(5):916–932, 2009.

[87] Avner Ottensooser, Alan Fekete, Hajo A Reijers, Jan Mendling, and Con Menictas. Making sense of business process descriptions: An experimental comparison of graphical and textual notations. *Journal of Systems and Software*, 85(3):596–606, 2012.

[88] Lukas Pilat and Hermann Kaindl. A knowledge management perspective of requirements engineering. In *Proceeding of the 5th International Conference on Research Challenges in Information Science*, pages 1–12. IEEE, 2011.

[89] Colin Potts. Software-engineering research revisited. 10(5):19–28, 1993.

[90] Helen C Purchase, Ray Welland, Matthew McGill, and Linda Colpoys. Comprehension of diagram syntax: an empirical study of entity relationship notations. *International Journal of Human-Computer Studies*, 61(2):187–203, 2004.

[91] Filippo Ricca, Massimiliano Di Penta, Marco Torchiano, Paolo Tonella, and Mariano Ceccato. The role of experience and ability in comprehension tasks supported by uml stereotypes. In *Proceedings of the 29th International Conference on Software Engineering*, volume 7, pages 375–384, 2007.

[92] Ioana Rus and Mikael Lindvall. Knowledge management in software engineering. *IEEE Software*, 19(3):26–38, 2002.

[93] Humphrey M Sabi, Faith-Michael E Uzoka, Kehbuma Langmia, and Felix N Njeh. Conceptualizing a model for adoption of cloud computing in education. *International Journal of Information Management*, 36(2):183–191, 2016.

[94] Johnny Saldaña. *The coding manual for qualitative researchers*. Sage, 2012.

[95] Faisal Saleh and Mohamed El-Attar. A scientific evaluation of the misuse case diagrams visual syntax. *Information and Software Technology*, 66:73–96, 2015.

[96] Riccardo Scandariato, Federica Paci, Katsiaryna Labunets, Koen Yskout, Fabio Massacci, Wouter Joosen, et al. Empirical assessment of security requirements and architecture: Lessons learned. In *Engineering Secure Future Internet Services and Systems*, pages 35–64. Springer, 2014.

[97] Riccardo Scandariato, Kim Wuyts, and Wouter Joosen. A descriptive study of microsoft's threat modeling technique. *Requirements Engineering*, pages 1–18, 2014.

[98] Riccardo Scandariato, Koen Yskout, Thomas Heyman, and Wouter Joosen. Architecting software with security patterns. *CW Reports*, 2008.

[99] Giuseppe Scanniello, Carmine Gravino, Marcela Genero, Jose'A Cruz-Lemus, and Genoveffa Tortora. On the impact of uml analysis models on source-code comprehensibility and modifiability. 23(2):13, 2014.

[100] Giuseppe Scanniello, Carmine Gravino, Michele Risi, Genoveffa Tortora, and Gabriella Dodero. Documenting design-pattern instances: A family of experiments on source-code comprehensibility. 24(3):14, 2015.

[101] Giuseppe Scanniello, Miroslaw Staron, Hakan Burden, and Rogardt Heldal. On the Effect of Using SysML Requirement Diagrams to Comprehend Requirements: Results from Two Controlled Experiments. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, pages 433–442, 2014.

[102] Ron Schmittling and Anthony Munns. Performing a security risk assessment. *ISACA Journal*, 1:18, 2010.

[103] Bruce Schneier. The importance of security engineering. *IEEE Security and Privacy*, (5):88, 2012.

[104] D.L. Schuirmann. On hypothesis-testing to determine if the mean of a normal-distribution is contained in a known interval. In *Biometrics*, volume 37, pages 617–617. International Biometric Soc, 1981.

[105] Ulrike Schultze and Dorothy E Leidner. Studying knowledge management in information systems research: discourses and theoretical assumptions. *Management Information Systems Quarterly*, pages 213–242, 2002.

[106] Ulrike Schultze and Charles Stabell. Knowing what you don't know? discourses and contradictions in knowledge management research. *Journal of Management Studies*, 41(4):549–573, 2004.

[107] Markus Schumacher, Eduardo Fernandez-Buglioni, Duane Hybertson, Frank Buschmann, and Peter Sommerlad. *Security Patterns: Integrating security and systems engineering*. John Wiley & Sons, 2013.

[108] Zohreh Sharafi, Alessandro Marchetto, Angelo Susi, Giuliano Antoniol, and Yann-Gaël Guéhéneuc. An empirical study on the efficiency of graphical vs. textual representations in requirements comprehension. In *Proceedings of the IEEE 21st International Conference on Program Comprehension*, pages 33–42. IEEE, 2013.

[109] Guttorm Sindre and Andreas L. Opdahl. Eliciting security requirements with misuse cases. *Requirements Engineering*, 10(1):34–44, 2005.

[110] Dag IK Sjøberg, Jo Erskine Hannay, Ove Hansen, Vigdis By Kampenes, Amela Karahasanovic, N-K Liborg, and Anette C Rekdal. A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering*, 31(9):733–753, 2005.

[111] Amina Souag, Raúl Mazo, Camille Salinesi, and Isabelle Comyn-Wattiau. Reusable knowledge in security requirements engineering: a systematic mapping study. *Requirements Engineering*, pages 1–33, 2015.

[112] Amina Souag, Camille Salinesi, Isabelle Wattiau, and Haris Mouratidis. Using security and domain ontologies for security requirements analysis. In *Proceeding of the 8th IEEE International Workshop on Security, Trust and Privacy for Software Application at the 37th IEEE International Computer Software and Applications Conference*, pages 101–107. IEEE, 2013.

[113] J. Michael Spector, M. David Merrill, Jan Elen, and M.J. Bishop. *Handbook of Research on Educational Communications and Technology*. Springer, 2014.

[114] Gary Stoneburner, Alice Goguen, and Alexis Feringa. Nist sp 800-30: Risk management guide for information technology systems. `http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf`, Last accessed: March 2016.

[115] Tor Stålhane and Guttorm Sindre. A comparison of two approaches to safety analysis based on use cases. In *Proceeding of the 26th International Conference on Conceptual Modeling*, volume 4801, pages 423–437, 2007.

[116] Tor Stålhane and Guttorm Sindre. Safety hazard identification by misuse cases: Experimental comparison of text and diagrams. In *Proceeding of the 9th International Conference on Model Driven Engineering Languages and Systems*, pages 721–735, 2008.

[117] Tor Stålhane and Guttorm Sindre. Identifying safety hazards: An experimental comparison of system diagrams and textual use cases. In *Proceeding of the 13th International Conference Enterprise, Business-Process and Information Systems Modeling*, volume 113, pages 378–392, 2012.

[118] Tor Stålhane and Guttorm Sindre. An experimental comparison of system diagrams and textual use cases for the identification of safety hazards. *International Journal of Information System Modeling and Design*, 5(1):1–24, 2014.

[119] Tor Stålhane, Guttorm Sindre, and Lydie Bousquet. Comparing safety analysis based on sequence diagrams and textual use cases. In *Proceeding of the 22nd International Conference on Advanced Information Systems Engineering*, volume 6051, pages 165–179, 2010.

[120] Anselm L. Strauss and Juliet M. Corbin. *Basics of qualitative research: techniques and procedures for developing grounded theory.* Sage Publications, 1998.

[121] Husam Suleiman and Davor Svetinovic. Evaluating the effectiveness of the security quality requirements engineering (square) method: a case study using smart grid advanced metering infrastructure. *Requirements Engineering*, pages 1–29, 2012.

[122] Mikael Svahnberg, Aybüke Aurum, and Claes Wohlin. Using students as subjects – an empirical evaluation. In *Proceeding of the 2nd International Symposium on Empirical Software Engineering and Measurement*, pages 288–290. ACM, IEEE, 2008.

[123] Stefan Taubenberger, Jan Jurjens, Yijun Yu, and Bashar Nuseibeh. Resolving vulnerability identification errors using security requirements on business process models. *Information Management and Computer Security*, 21(3):202–223, 2013.

[124] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. User acceptance of information technology: Toward a unified view. *Management Information Systems Quarterly*, pages 425–478, 2003.

[125] Iris Vessey. Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, 22(2):219–240, 1991.

[126] Rodolfo Villarroel, Eduardo Fernández-Medina, and Mario Piattini. Secure information systems development – a survey and comparison. *Computers and Security*, 24(4):308–321, 2005.

[127] M. McLure Wasko and Samer Faraj. "It is what one does": why people participate and help others in electronic communities of practice. *Journal of Strategic Information Systems*, 9(2):155–173, 2000.

[128] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Bjöorn Regnell, and Anders Wesslén. *Experimentation in software engineering*. Springer, 2012.

[129] Idan Wolf and Pnina Soffer. Supporting BPMN model creation with routing patterns. In *Proceedings of the 26th International Conference on Advanced Information Systems Engineering*, pages 171–181. Springer, 2014.

[130] Robert E Wood. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes*, 37(1):60–82, 1986.

[131] Kim Wuyts, Riccardo Scandariato, and Wouter Joosen. Empirical evaluation of a privacy-focused threat modeling methodology. *Journal of Systems and Software*, 96:122–138, 2014.

[132] Robert K Yin. *Qualitative research from start to finish*. Guilford Press, 2010.

[133] Koen Yskout, Riccardo Scandariato, and Wouter Joosen. Do security patterns really help designers? In *Proceedings of the 37th International Conference on Software Engineering*, pages 292–302. IEEE, 2015.

[134] Cheng Zhang and David Budgen. What do we know about the effectiveness of software design patterns? *IEEE Transactions on Software Engineering*, 38(5):1213–1231, 2012.