Empir Software Eng (2017) 22:3017–3056
DOI 10.1007/s10664-017-9502-8

Model comprehension for security risk assessment: an empirical comparison of tabular vs. graphical representations

Katsiaryna Labunets[1] · Fabio Massacci[1] · Federica Paci[2] · Sabrina Marczak[3] · Flávio Moreira de Oliveira[3]

Published online: 24 February 2017

Abstract Tabular and graphical representations are used to communicate security risk assessments 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 relationships should be better captured by graphs. In this paper we report the results of two studies performed in two countries with 69 and 83 participants respectively, in which we assessed the effectiveness of tabular and graphical representations with respect to extraction correct information about security risks. The experimental results show that tabular risk models are more effective than the graphical ones with respect to simple comprehension tasks and in some cases are more effective for complex comprehension tasks. We explain our findings by proposing a simple extension of Vessey's cognitive fit theory as some linear spatial relationships could be also captured by tabular models.

Communicated by: Andrea De Lucia

✉ Katsiaryna Labunets
katsiaryna.labunets@unitn.it

Fabio Massacci
fabio.massacci@unitn.it

Federica Paci
F.M.Paci@soton.ac.uk

Sabrina Marczak
sabrina.marczak@pucrs.br

Flávio Moreira de Oliveira
flavio.oliveira@pucrs.br

[1] University of Trento, Trento, Italy

[2] University of Southampton, Southampton, UK

[3] Pontifícia Universidade Catòlica do Rio Grande do Sul (PUCRS) University, Porto Alegre, Brazil

Keywords Empirical study · Security risk assessment · Risk modeling · Comprehensibility · Cognitive fit

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" (Mead et al. 2004). 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., fit 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." (Landoll and Landoll 2005, 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 (Labunets et al. 2014a, 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 (Mellado et al. 2006)) whereas academic based methods use *graphical modeling notations* (eg. SI^* (Giorgini et al. 2005), Secure Tropos (Mouratidis and Giorgini 2007), ISSRM (Matulevičius et al. 2008), or CORAS (Lund et al. 2011)). Yet, there is limited empirical evidence whether one of the two risk modeling notation better supports the comprehension of security risks. Hence, this paper aims to investigate the following research questions:

$RQ1$ *Which risk modeling notation, tabular or graphical, is more effective in extracting correct information about security risks?*

$RQ2$ *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 (Wood 1986). 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.

In the rest of the paper we discuss related work Section 2, describe the study design Section 3, and report the experiments realization Section 4. Section 5 presents the results of the analysis and Section 6 discusses their implications. Finally we discuss the threats to validity of our study Section 7 and conclude the paper Section 8.

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 (Vessey 1991; Moody 2009); other studies have compared different notations from a conceptual point of view (Kaczmarek et al. 2015; Saleh and El-Attar 2015). Several empirical studies have compared graphical and textual representations for requirements (Sharafi et al. 2013; Stålhane and Sindre 2008; Stålhane et al. 2010; Stålhane and Sindre 2014), software architectures (Heijstek et al. 2011), and business processes (Ottensooser et al. 2012). Studies that focus on comparing textual and visual notations for security risk models are less frequent (Hogganvik and Stolen 2005; Grondahl and Lund 2011) or compared the effectiveness of tabular or graphical methodologies as whole (Massacci and Paci 2012; Labunets et al. 2013, 2014b) as opposed to the specific aspect of comprehensibility.

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. 2014) 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. (2014a) 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. (2013) 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. (2013) 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. (2011) 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. (2012) 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.

2.2 Empirical Comparisons of Security Modeling Notations

In the specific domain of modeling security issues, Stalhane et al. conducted a series of experiments (Stålhane and Sindre 2008; Stålhane et al. 2010; Stålhane and Sindre 2014) to compare the effectiveness of textual and visual notations in identifying safety hazards during security requirements analysis. Stålhane and Sindre (2008) 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 (Stålhane et al. 2010; Stålhane and Sindre 2012; Stålhane and Sindre 2014) 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. The first work, Hogganvik and Stolen (2005) 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. The second work, Grondahl and Lund (2011) 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 (Labunets et al. 2013, 2014b). 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. (2013) reported that "some of the participants indicated that a visual representation for threat would be better that a tabular one", and in (Labunets et al. 2014b) 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 (Labunets et al. 2014b, Table III). In this paper we explore whether such preference may be explained by the widely held belief that graphical representations are easier to read.

3 Study Planning

3.1 Motivation

In our previous study (Labunets et al. 2014a) 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 (Labunets et al. 2014a, 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."

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 (Hadar et al. 2013; Scanniello et al. 2014b) and data modeling (De Lucia et al. 2010; Purchase et al. 2004). 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.

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 (1986), 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" (Wood 1986, p. 65);
- "The required acts for the creation of a defined product [output] can be described at any one of several levels of abstraction..." (Wood 1986, 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..." (Wood 1986, 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:

$$QC_i = |IC_i| + |R_i| + |A_i|, \tag{1}$$

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 (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 (Hadar et al. 2013; Agarwal et al. 1999). Table 15 in Appendix A reports the post-task questionnaire that we proposed to our participants.

3.4 Selection of Risk Modeling Notations

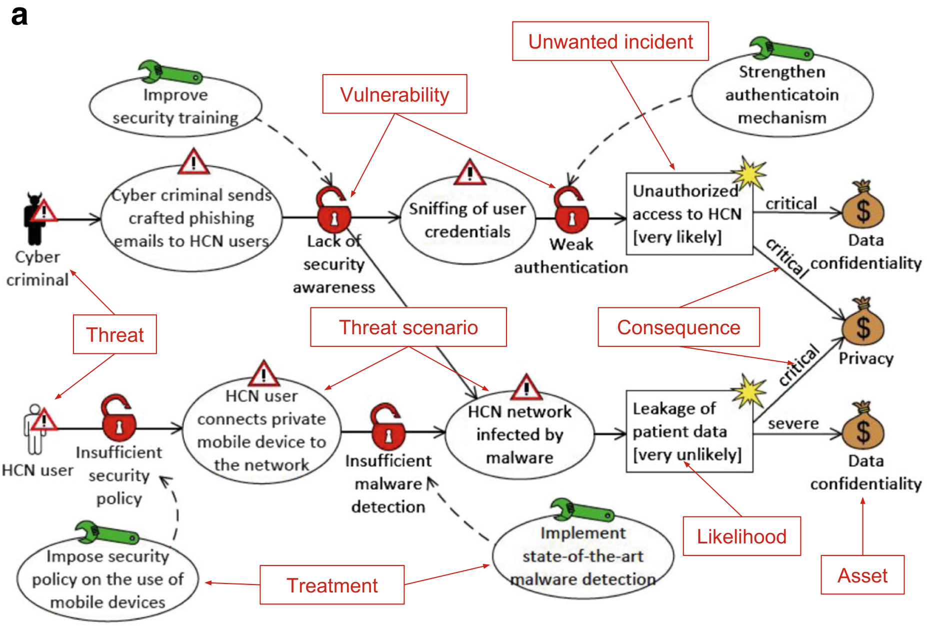
There are many different methods for security requirements engineering and risk assessment that use either graphical, or tabular, or mix of two representations. To make the study fair and representative we need to find notations that have similar level of expressiveness and cover the core security concepts 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. (2010) 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. (2007) (see Table 3 in (Fabian et al. 2010)). The core concepts that emerged from the studies are *asset, threat, vulnerability, risk, and security control.*

The comparison by Fabian et al. (2010) showed that only several methods adopted these concepts, namely tabular *SREP* (Mellado et al. 2006), graphical *CORAS* (Lund et al. 2011), and model-based *information system security risk management (ISSRM)* approach proposed by Mayer et al. (2005). The ISSMR method initially used i* models to support risk analysis and has been later adapted to by Matulevičius et al. (2008) to combine the graphical-based method proposed by Mouratidis and Giorgini (2007) *Secure Tropos*.

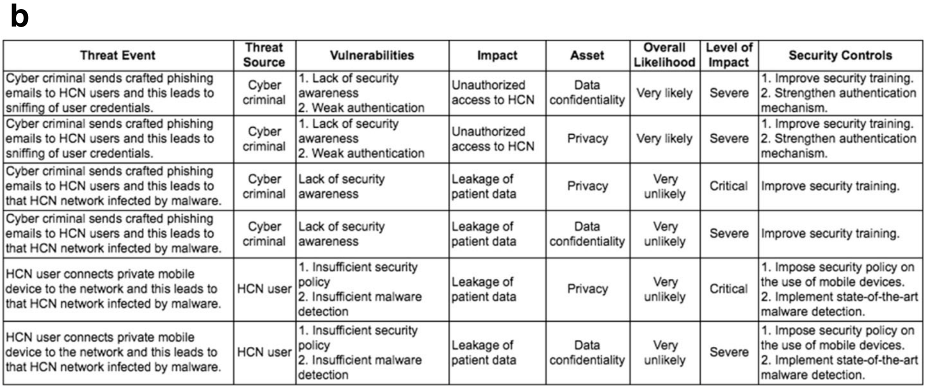
To the best of our knowledge, the work by Massacci and Paci (2012) is the only study that empirically investigated and compared different security methods including Secure Tropos, CORAS, $si*$, and Security Argumentation. Both CORAS and Secure Tropos methods were empirically evaluated by Massacci and Paci (2012). 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 (Stoneburner et al. 2002) 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


a
Unwanted incident
Improve security training
Vulnerability
Strengthen authenticatoin mechanism
Cyber criminal sends crafted phishing emails to HCN users
Sniffing of user credentials
Unauthorized access to HCN [very likely]
critical
Data confidentiality
Cyber criminal
Lack of security awareness
Weak authentication
critical
Threat
Threat scenario
Consequence
Privacy
critical
HCN user connects private mobile device to the network
HCN network infected by malware
Leakage of patient data [very unlikely]
severe
HCN user
Insufficient security policy
Insufficient malware detection
Data confidentiality
Implement state-of-the-art malware detection
Likelihood
Asset
Impose security policy on the use of mobile devices
Treatment

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.

NIST table row entries

Fig. 1 Fragment of a risk model in graphical and tabular notations

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 usually documented in different tables. Figure 1a shows an example of CORAS treatment diagram related to the risk of a Healthcare Collaborative Network, and Fig. 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 Stolen (2005) 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 7 and 8 in Appendix A illustrate the full graphical and tabular risk models that we provided to our participants.

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 (Agarwal et al. 1999; Hadar et al. 2013; Scanniello et al. 2014b, 2015): *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:

$$precision_{m,s,q} = \frac{|answer_{m,s,q} \cap correct_q|}{|answer_{m,s,q}|}, \tag{2}$$

$$recall_{m,s,q} = \frac{|answer_{m,s,q} \cap correct_q|}{|correct_q|}, \tag{3}$$

$$F_{m,s,q} = 2 * \frac{precision_{m,s,q} \times recall_{m,s,i}}{precision_{m,s,q} + recall_{m,s,q}}, \tag{4}$$

$$F_{m,s} = \text{mean}(\cup_{q \in \{1 \ldots N_{questions}\}} F_{m,s,q}\} \tag{5}$$

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.

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 (Abrahao et al. 2013) and aggregate all answers to calculate precision and recall for the individual participant.

$$precision_{m,s} = \frac{\sum_{q=1}^{N_{questions}} |answer_{m,s,q} \cap correct_q|}{\sum_{q=1}^{N_{questions}} |answer_{m,s,q}|}, \tag{6}$$

$$recall_{m,s} = \frac{\sum_{q=1}^{N_{questions}} |answer_{m,s,q} \cap correct_q|}{\sum_{q=1}^{N_{questions}} |correct_q|}, \tag{7}$$

$$F_{m,s} = 2 * \frac{precision_{m,s} \times recall_{m,s}}{precision_{m,s} + recall_{m,s}}. \tag{8}$$

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

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 ($RQ1$). 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 2), and therefore, we did not make any assumptions in this regard. For example, Stålhane and Sindre (2014) and Hogganvik and Stolen (2005) 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 1.

3.7 Experimental Design

In the first study we chose a *between-subject design* with one factor (risk modeling notation) and two treatments (graphical and tabular risk models) to avoid interference between the treatments (MacKenzie 2012, 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

Table 1 Experimental Hypotheses

Hyp	Null Hypothesis	Alternative Hypothesis
$H1$	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
$H2$	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

Table 2 Experimental design of the second study

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

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)

are presented in Appendix A in Figs. 7 and 8 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 (Hadar et al. 2013; De Lucia et al. 2010; Hoisl et al. 2014), for the second study we chose a *within-participants 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 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]

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 5.4

[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

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 $RQ2$. Table 3 presents the distribution of the questions by the number of information cues, relationships and judgments present in the question. Table 17 in Appendix A 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.

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 16 and 17 in Appendix A report the comprehension questionnaire for the graphical risk model for both studies. The

Table 3 Comprehension questionnaire design

	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

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 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.

3.9 Analysis Procedure

We test the null hypothesis $H1_0$ using an unpaired statistical test in the first study as we have a between-participants design, and a paired statistical test in the second study because of a within-participants 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 Mann–Whitney (MW) and Wilcoxon tests respectively.

We investigate the effect of task complexity and test the null hypothesis $H2_0$ 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 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 and visualize the results we used `RStudio`[3] with the following packages:

- Package "car" by (Fox and Weisberg 2011) for Levene's test for homogeneity of variance (function *leveneTest*),
- Package "stats" by (R Core Team 2016) for Shapiro-Wilk normality test (function *shapiro.test*),
- Package "exactRankTests" by (Hothorn and Hornik 2015) for Wilcoxon and Mann-Whitney tests (function *wilcox.exact*). We use it because this package can handle tied observations that present in our samples.
- To produce graphics we used the combination of the following packages: "ggplot2" by (Wickham 2009), "gtable" by (Wickham 2016), and "grid" by (R Core Team 2016).

4 Study Realization

4.1 Experiments Execution

Table 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

[3] www.rstudio.com

Table 4 Participants Distribution to Treatments - study 1

In total 36 participants completed the comprehension task using the graphical risk model and 32 participants used the tabular notation

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

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 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) 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.

4.2 Demographics

Table 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

[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 5 Participants Distribution to Treatments - study 2

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

In total we had 83 participants who were randomly assigned to one of four groups. The description of the groups see in Table 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

other graphical modeling notations. The participants only had very basic knowledge of the application scenario.

Table 7 summarizes the demographic information about the participants of our experiments for the second study. Most participants (51 %) reported that they had working experience. The participants of the second study had slightly better security knowledge and slightly worse knowledge of modeling languages compared to the participants of the first study (see Table 6). They also had very basic knowledge of the application scenarios.

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 $RQ1$ we proceeded with a non-parametric MW test for the results of the first study as it has between-subject design and with Wilcoxon test for the second study because it has within-subject design. In $RQ2$ 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.

Table 6 Demographic statistics - study 1

Variable	Scale	Mean/Med.	Distribution
Age	Years	25.8 (mean)	45 % were 19–23 yrs old; 36 % 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	0–4 (Novice–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

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

Table 7 Demographic statistics - study 2

Variable	Scale	Mean/Med.	Distribution
Age	Years	26.4 (mean)	25.3 % were 21–23 yrs old; 55.4 % were 24–29 yrs old; 19.3 % 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); 15 % 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	0–4 (Novice–Expert)	1 (median)	19 % novices; 52 % beginners; 18 % competent users; 6 % proficient; 5 % experts
Expertise in modeling languages	0–4	2 (median)	16 % novices; 33 % beginners; 36 % competent users; 13 % proficient users; 2 % 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)	81 % novices; 18 % beginners; 1 % experts

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

5.1 RQ1: Effect of Risk Modeling Notation on Comprehension

Tables 8 and 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 18 and 19 in Appendix.

Table 8 Descriptive statistics of precision and recall by modeling notation - study 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						
1. UNITN-MCS	0.90	0.92	0.06	0.84	0.88	0.11
1. PUCRS-MCS	0.82	0.87	0.12	0.70	0.74	0.10
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						
1. UNITN-MCS	0.89	0.89	0.07	0.75	0.78	0.15
1. PUCRS-MCS	0.89	0.93	0.09	0.61	0.66	0.11
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 9 Descriptive statistics of precision and recall by modeling notation - study 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						
2. POSTE	0.92	0.96	0.09	0.80	0.88	0.19
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						
2. POSTE	0.87	0.88	0.11	0.64	0.65	0.19
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

Figure 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 8 and 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 $H1_0$ is tested with Wilcoxon and MW tests and the results presented in Table 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.

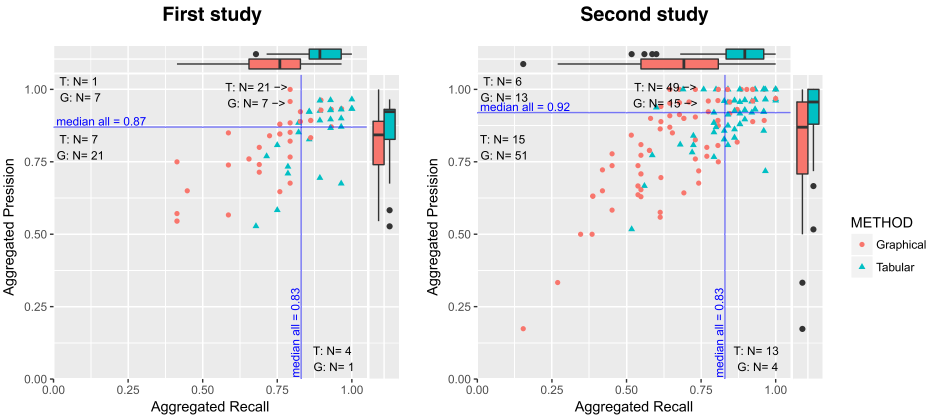


Fig. 2 Distribution of participants' precision and recall by modeling notation

Table 10 RQ1 – Summary of Experimental Results by Modeling Notation

	Experiment	#part.	#obs.	μ_T-μ_G	σ	p-value	Cohen's	d
Precision	1. UNITN-MCS	35	35	0.06	0.12	0.024	Small	0.49
	1. PUCRS-MCS	13	13	0.13	0.18	0.046	Medium	0.71
	1. PUCRS-BSC	21	21	0.01	0.22	0.66	Negligible	0.06
	2. POSTE	41	82	0.12	0.15	$6.7 \cdot 10^{-5}$	Medium	0.79
	2. UNITN	42	84	0.09	0.12	$4.1 \cdot 10^{-6}$	Large	0.81
	Study 1: Overall	69	69	0.06	0.17	0.018	Small	0.32
	Study 2: Overall	83	166	0.11	0.13	$1.9 \cdot 10^{-8}$	Medium	0.79
Recall	1. UNITN-MCS	35	35	0.14	0.14	0.002	Large	0.95
	1. PUCRS-MCS	13	13	0.27	0.15	0.001	Very large	1.87
	1. PUCRS-BSC	21	21	0.15	0.21	0.054	Medium	0.70
	2. POSTE	41	82	0.23	0.16	$1.9 \cdot 10^{-9}$	Very large	1.46
	2. UNITN	42	84	0.18	0.14	$5.7 \cdot 10^{-9}$	Very large	1.25
	Study 1: Overall	69	69	0.16	0.17	$5.0 \cdot 10^{-6}$	Large	0.98
	Study 2: Overall	83	166	0.20	0.15	$4.1 \cdot 10^{-13}$	Very large	1.35

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

5.2 RQ2: Effect of Task Complexity on Comprehension

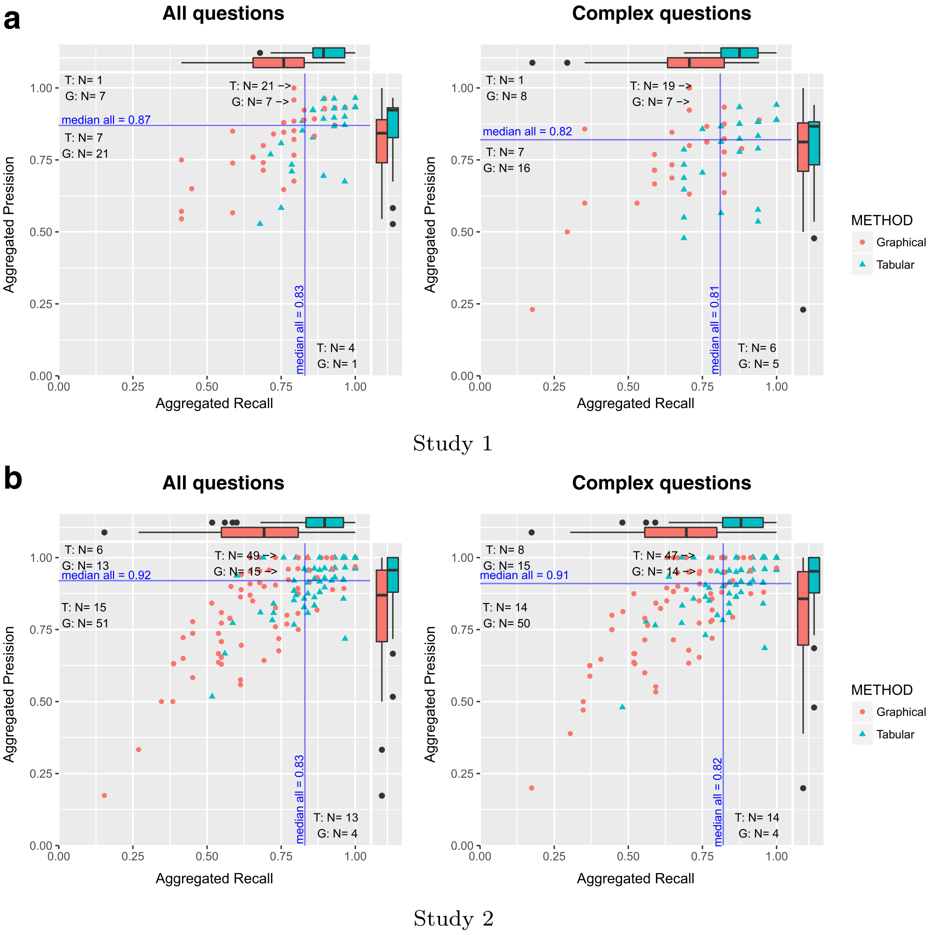
Figure 3a and b compare the distribution of precision and recall of the participants' responses to full comprehension task (Q1–Q12) (left) and only to the complex questions (right), namely question complexity level > 2.

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 the overall median recall threshold in tabular and graphical group respectively.

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. Figure 4a and b shows that there is no significant interaction between precision, recall and modeling notation.

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 inprecision between simple and complex questions. As there is no major interaction between risk model notation and either



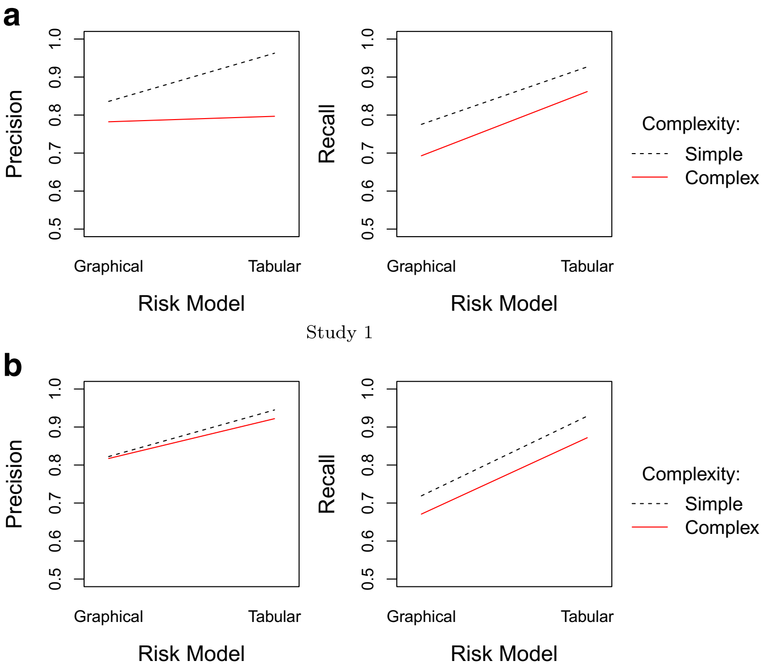
Study 1

Study 2

Fig. 3 Distribution of participants' precision and recall by task complexity

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 the second research question. 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 18 and 19 in Appendix). The formulation is essentially identical to (5) except that q only ranges over the questions with complexity ℓ.

Tables 11 and 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 participants of experiments PUCRS-MCS in the first 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.



Study 1

Study 2

Fig. 4 Interaction among risk modeling notation and task complexity

The $H2_0$ is tested with Wilcoxon test and the results reported in Table 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.

Table 11 Descriptive statistics of F-measure by task complexity - study 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	1. UNITN-MCS	0.98	1.00	0.04	0.83	0.88	0.10
	1. PUCRS-MCS	0.90	1.00	0.17	0.82	0.86	0.13
	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	1. UNITN-MCS	0.85	0.86	0.15	0.75	0.80	0.17
	1. PUCRS-MCS	0.67	0.66	0.18	0.64	0.65	0.13
	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 12 Descriptive statistics of F-measure by task complexity - study 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	2. POSTE	0.93	1.00	0.20	0.89	0.90	0.09
	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.	2. POSTE	0.76	0.86	0.27	0.70	0.75	0.18
	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

In Appendix B 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.

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. Table 15a and b 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).

Both for tabular and graphical risk 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

Table 13 RQ2 – Summary of Experimental Results by Tasks' Complexity

	Experiment	#part.	#obs.	μ_C-μ_S	σ	p-value	Cohen's	d
Tabular	1. UNITN-MCS	17	17	−0.14	0.08	$1.5 \cdot 10^{-5}$	Very large	1.69
	1. PUCRS-MCS	7	7	−0.08	0.23	0.30	Small	0.36
	1. PUCRS-BSC	9	9	−0.10	0.23	0.055	Small	0.45
	2. POSTE	41	41	−0.04	0.24	0.0003	Negligible	0.19
	2. UNITN	42	42	−0.04	0.20	0.002	Negligible	0.18
	Study 1: Overall	33	33	−0.12	0.17	$1.1 \cdot 10^{-5}$	Medium	0.68
	Study 2: Overall	83	83	−0.04	0.22	$6.4 \cdot 10^{-6}$	Negligible	0.18
Graphical	1. UNITN-MCS	18	18	−0.09	0.23	0.03	Small	0.41
	1. PUCRS-MCS	6	6	−0.03	0.25	1.00	Negligible	0.11
	1. PUCRS-BSC	12	12	−0.07	0.30	0.15	Small	0.23
	2. POSTE	41	41	−0.06	0.36	0.08	Negligible	0.16
	2. UNITN	42	42	0.00	0.33	0.41	Negligible	-0.00
	Study 1: Overall	36	36	−0.07	0.27	0.01	Small	0.28
	Study 2: Overall	83	83	−0.03	0.35	0.06	Negligible	0.08

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 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.

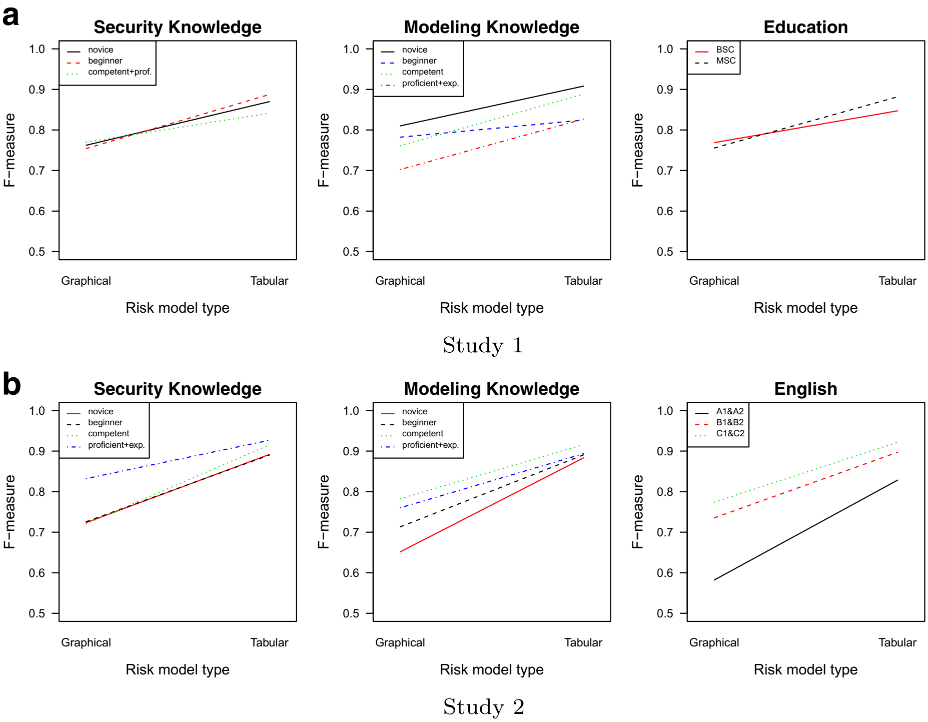
5.4 Co-factor Analysis

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 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 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 2 We investigated a possible learning effect that may be caused by between-participants design. Figure 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



Study 1

Study 2

Fig. 5 Interaction of modeling notations with expertise co-factors

$p = 0.88$ for the scenario and risk modeling notation and $p = 0.96$ for the session and risk model type.

6 Discussion and Implications

In this section we discuss our results with respect to the hypotheses presented in Section 3.6. We also discuss possible explanation of the outcomes and their implications to research and practice.

The first null hypothesis $H1_0$ (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 $H2_0$ (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.

In summary, participants who applied the tabular risk model gave more precise and complete answers to the comprehension questions when requested to find simple and complex information about threats, vulnerabilities, or other elements of risk models. Further, participants showed equal preference when asked about their own perception of the two risk modeling notations (Q5 and Q6 in post-task questionnaire). These results are consistent across both our studies (See Table 14).

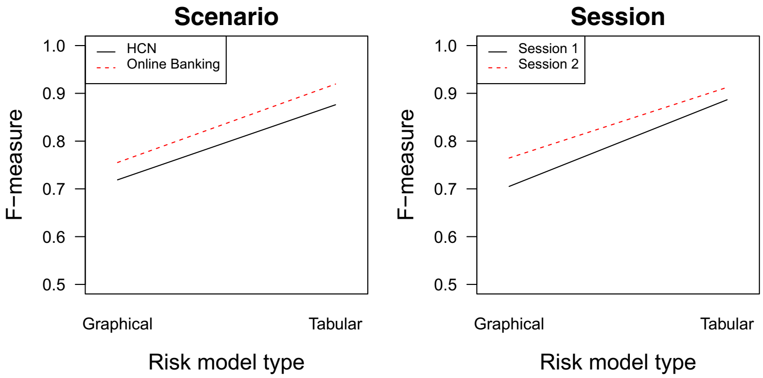


Fig. 6 Interaction of scenario and session vs modeling notation - study 2

Such result is partly surprising as the theory of cognitive fit (Vessey 1991) suggests that a match between problem representation and task should result in a better problem solving performance. As our questions are all about finding relations between elements, a graphical notation with its explicit representation of spatial relations should have a better comprehension performance. It was also against the original expectation of the authors' team who invented one of the most cited graphical method for security requirements engineering (Giorgini et al. 2005). Indeed, in all our own previous studies where graphical and textual risk assessment methods were compared (Labunets et al. 2013, 2014b; Massacci and Paci 2012), participants systematically perceived the graphical risk assessment method (the some one used in this study) as superior in terms of ease of use and effectiveness (albeit they often had the same actual effectiveness).

We argue that such difference in comprehension between the risk modeling notation can be explained by cognitive fit theory itself if we do not unnecessarily restrict spatial relationships to graphs as initially argued by Vessey (1991). Whereas columns are clearly devised for looking up elements, *tables implicitly capture elementary linear spatial relationships by their rows*: each row relates some column elements to each other.

Consider again our example question 'What is the highest possible consequence for the asset "Data confidentiality" that Cyber criminal can cause?'. Finding the first consequence in Fig. 1a requires walking one straight line from left to right. Most relations in the models and our natural questions are linear or tree relations.

The same left-to-right eye's flight can be performed in the Tabular representation in Fig. 1b after finding the first instance of "Cyber criminal" in the appropriate column. This example illustrates that the row itself captures the linear relationship. Therefore, according to cognitive fit theory both representations would be equally well suited for the task (of finding one consequence).

However, our question is *not* about finding one consequence, is about finding the most critical consequence which is the important question to ask given the role of risk analysis to prioritize countermeasures (see the experts' opinion discussed in (Labunets et al. 2014a)).

Graphical notation's ability to "summarize" elements (there is only one single "Cyber criminal") and its a minimal duplication (such as reporting twice the 'Data confidentiality' asset to avoid cluttering the diagram) should make it easier to cluster the elements and therefore to report more consequences. In contrast, our example Table in Fig. 1b has four instances of "Cyber criminal"s and three instances of "Data confidentiality". More elements to search for would therefore mean a higher likelihood to omit one element give the limited time.

Table 14 Post-task questionnaire results

	Tabular			Graphical		
Q#	Mean	Median	sd	Mean	Median	sd
(a) Study 1						
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 2						
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%)		

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)

As apparent from our experimental results, finding the second consequence turns out to be harder when using a graphical notation: it requires to navigate through the graph in a tree structure: right first, then down through the node “HCN network infected by malware” and then again right to the end. The analysis of the spatial relationships in a table can be seen as a sequence of look ups (on which tabular methods are notoriously good at) followed by a quick spatial relationship analysis (by row). Therefore, the higher number of look ups is apparently compensated by the easier processing of linear spatial relations against tree-based relationships.

This theory could be tested by performing additional experiments in which progressively more complex questions are asked to determine whether a sweet spot exists where graphical models would be identical or easier to understand than tabular models. When questions could no longer be subsumed by sequences of look ups and linear relations the performance of the graphical notation should be superior. Yet, if the models were to get too large for such questions to make sense, then both tabular and graphical models would produce poor results.

Implications for practice Translating Table 10 into practical values, participants exposed to a risk analysis represented with a graphical notation gave one wrong answer out of ten and failed to report one key element out of five more than participants exposed to a tabular risk modeling notation. Given the role of risk analysis such failures may be considered unacceptably high for some domain.

The adoption of a tabular notation by international standards might have been dictated by simplicity considerations but turns out to be better from comprehension purposes. 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 provide a notation closer to natural language. The stakeholders also may benefit from using the 'look-up' bonus of tables with filters and sorting option in the tables.

It is however unclear whether tabular notation might scale to very large risk assessments as our result in Table 13 showed that there is a drop in effectiveness when faced with more complex questions. Such drop is small (less than one question out of ten is answered incorrectly) but is nonetheless significant. In contrast, graphical models did not suffer from such drop albeit it might just be because their performance was already low. More investigations are needed in this respect.

Another practical issue to investigate is the impact on comprehension of the use of a standard terminology for the definition of the specific instances of threat and controls (as opposed to the class names of the elements). Security catalogues are widely used in industry and they have a notable effect on the production of risk assessments (see (De Gramatica et al. 2015)). The use of a well defined terminology might also ease the comprehension tasks, especially for complex questions.

Implications for research 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.

The experimental results showed that tabular notation is more effective than the graphical one in extracting correct information about security risks and we have argued that such performance might be due to the ability of a tabular notation to capture simple linear relationships. As we have discussed above such theory could be tested by further experiments where either questions or models are increasingly made more complex. There should be a point when either both notations perform poorly or the tabular notation ability to capture simple linear relations can no longer cope with complex relationships captured by a graphical notation.

Also task complexity factor requires further investigation. 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. Only for Judgements there seems to be a significant drop in comprehension for both graphical and tabular modeling notations (see Fig. 11). Therefore, task complexity should be always taken into account when researchers investigate the comprehensibility of different representations.

Indeed, the apparent contradiction between this result and our own previous research that we mentioned above (Labunets et al. 2013, 2014b; Massacci and Paci 2012) could be well explained by the difference in task complexity: in those studies participants had to *produce* models in the required notation. These studies were full-scale applications of security risk assessment methods to real-sized application scenarios that lasted for several weeks.

The generalization of our results is of course limited by our experimental set-up and we discuss the threats to validity more in details in the next section. We do not believe that different experiments would produce different results by changing the tabular notation as almost all standards are based on very similar tables. Yet there might be other modeling graphical notations that could perform better. From this perspective, the past experiments reported in (Grondahl and Lund 2011; Massacci and Paci 2012) give us confidence that we have selected one of the best graphical risk analysis notation available at the time of writing.

Beyond comprehension: retention Another, orthogonal avenue of research goes beyond the simple task of comprehension and we would like to thank our reviewer for pointing out this possibility. We term the phenomenon *retention* retention as it is a ‘Gestalt’ memorization of the key aspect of the risk assessment *after* the assessment has been presented and is no longer available.

The experiment could use the existing comprehension tasks reported in this paper but the models will be provided to participants only for a limited time to read and memorize. Then participants have to answer the questions without having the models available.

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 (Hadar et al. 2013; Ricca et al. 2007). However, the design of the questionnaire may be strongly favoring one treatment over the other. Inspired by similar studies (Heijstek et al. 2011; Sharafi et al. 2013), 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 Fig. 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 15a and b). 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 (Dunning et al. 2003), 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 Fig. 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 15a and b, most of the participants clearly understood the objectives of the study and the task to be performed.

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 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 (Kabacoff 2015).

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 (Svahnberg et al. 2008) 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 ((Stoneburner et al. 2002, Section 2.1) or (BSI 2012, 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).

8 Conclusion and Future Work

This paper has reported the results from a replication of experiments aimed at investigating the actual comprehension of a security risk represented using tabular and graphical modeling notations. In particular, the experimentation consisted of two studies of three and two replicated experiments, involving undergraduate students (21), master students (90) and graduate students in a professional master (41), in several different locations. The comprehension task was reading a risk model in either the tabular and graphical notation and answering questions on the model.

The results showed that tabular risk models are more effective than graphical ones with respect to extracting relevant information about security risks. The effect is medium in terms of precisions of their answers and large in terms of recall. We believe that these results can be explained by a simple extension of Vessey’s cognitive fit theory (Vessey 1991) as some linear spatial relationships can also (and possibly more easily) be captured by tabular models. Hence, for some natural comprehension questions about relationships among elements in a model the tabular representation also has a good fit in terms of matching tasks with representation.

The experiments provided less evidence on the impact on task complexity as defined by Wood (1986) and adapted by us to the comprehension of risk models in terms of questions involving different information cues, different relationships and different judgements. Only for participants using the tabular modeling notation there is a small drop in the level of comprehension as assessed by the F-measure of their answers.

The dataset with the results of reported experiments is openly available for research purposes[5] and the additional material for replication can be found on the web page of our research group.[6]

We plan to replicate our study with security professionals, as well as investigating further the effect of the modeling notation on the *retention* of the key information of a risk assessment i.e on the precision and recall of participants' answers *after* the model has been presented to them but is not readily available for consultation. The empirical findings would have major implications for practice as we can expect that most people would only refer back to the actual risk analysis on occasional basis.

Acknowledgements This work has been partly supported by the SESAR JU WPE under contract 12-120610-C12 (EMFASE). We would like to thank B. Solhaug and K. Stølen from SINTEF for support in the definition of the CORAS models. L. Allodi helped us to organized the first experiment for the second study in Cosenza.

Appendix A: Additional Data

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.

Fig. 7 Risk Model for HCN Scenario in Tabular Notation Provided to the Subjects

[5] Dataset https://securitylab.disi.unitn.it/doku.php?id=datasets.

[6] Materials for experiment replication https://securitylab.disi.unitn.it/doku.php?id=unitn-comprehensibility-exp-2015.

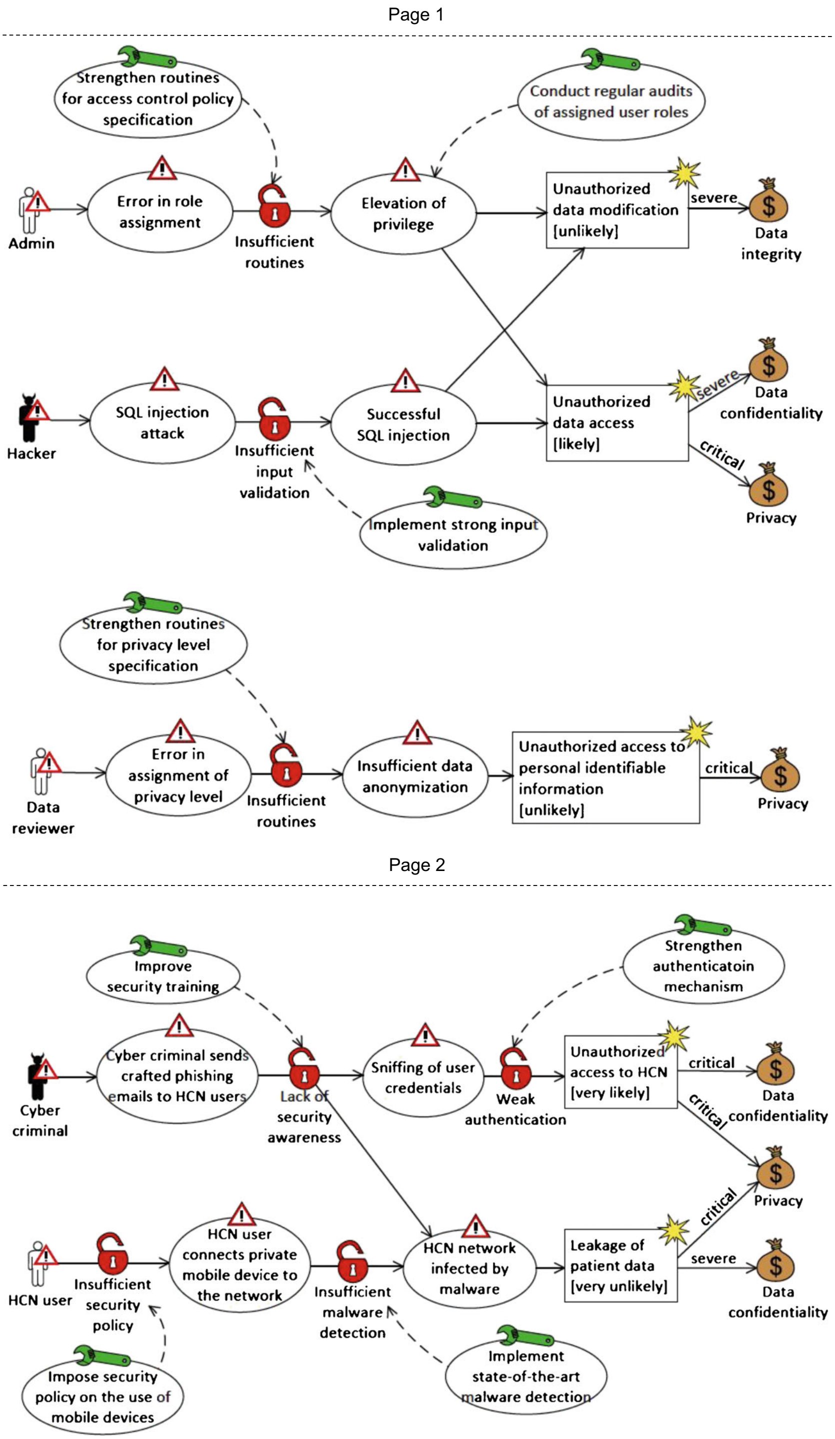

Page 1
Strengthen routines for access control policy specification
Conduct regular audits of assigned user roles
Admin
Error in role assignment
Insufficient routines
Elevation of privilege
Unauthorized data modification [unlikely]
severe
Data integrity
Hacker
SQL injection attack
Insufficient input validation
Successful SQL injection
Unauthorized data access [likely]
severe
Data confidentiality
critical
Privacy
Implement strong input validation
Strengthen routines for privacy level specification
Data reviewer
Error in assignment of privacy level
Insufficient routines
Insufficient data anonymization
Unauthorized access to personal identifiable information [unlikely]
critical
Privacy
Page 2
Improve security training
Strengthen authenticatoin mechanism
Cyber criminal
Cyber criminal sends crafted phishing emails to HCN users
Lack of security awareness
Sniffing of user credentials
Weak authentication
Unauthorized access to HCN [very likely]
critical
Data confidentiality
critical
Privacy
critical
HCN user
Insufficient security policy
HCN user connects private mobile device to the network
Insufficient malware detection
HCN network infected by malware
Leakage of patient data [very unlikely]
severe
Data confidentiality
Impose security policy on the use of mobile devices
Implement state-of-the-art malware detection

Fig. 8 Risk Model for HCN Scenario in Graphical Notation Provided to the Subjects

Table 15 Post-Task Questionnaire

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?

This is the post-task questionnaire that we distributed to the subjects. 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

Table 16 Comprehension Questions for Graphical Risk Model (Study 1)

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:

Table 16 (continued)

Q#	Complexity	Question statement
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:

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

Table 17 Comprehension Questions for Graphical Risk Model (Study 2)

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.

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

Table 18 Precision and recall by questions, study 1

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	0.58	1.00	0.50	36	0	0.42	0.00	0.50
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	0.28	36	0	0.80	1.00	0.37
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	0.50	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	0.58	1.00	0.50	36	0	0.42	0.00	0.50
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	0.88	1.00	0.27	36	0	0.73	1.00	0.37

The most significant difference (≥ 0.2) in precision was observed for Q1, Q6 and in recall for Q2, Q6-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

Table 19 Precision and recall by questions, study 2

Q#	Comp-	Tabular					Graphical				
	lexity	#obs.	∅	mean	med.	sd	#obs.	∅	mean	med.	sd
Precision											
Q1	2	83	1	0.94	1.00	0.24	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

Table 19 (continued)

Q#	Comp-	Tabular					Graphical				
	lexity	#obs.	∅	mean	med.	sd	#obs.	∅	mean	med.	sd
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	0.72	1.00	0.45
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	0.27	83	4	0.80	1.00	0.39
Recall											
Q1	2	83	1	0.94	1.00	0.24	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	0.72	1.00	0.45
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

The most significant difference ($\geq$ 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

Appendix B: Effect of Task Complexity Components on the Risk Model Comprehension

The following figures show the interaction plots between F-measure by model type (graphical vs. tabular) and the levels of IC (Fig. 9), R (Fig. 10), and the presence of the judgment component (Fig. 11).

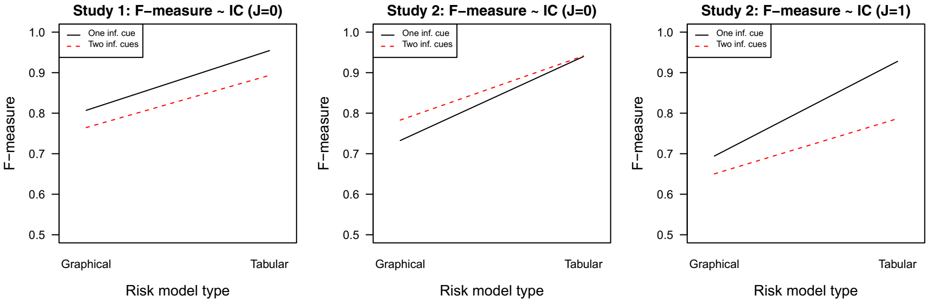


Fig. 9 Effect of complexity (IC) on F-measure

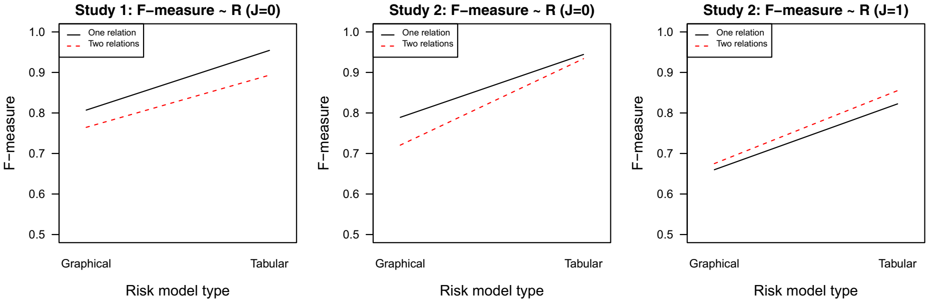


Fig. 10 Effect of complexity (R) on F-measure

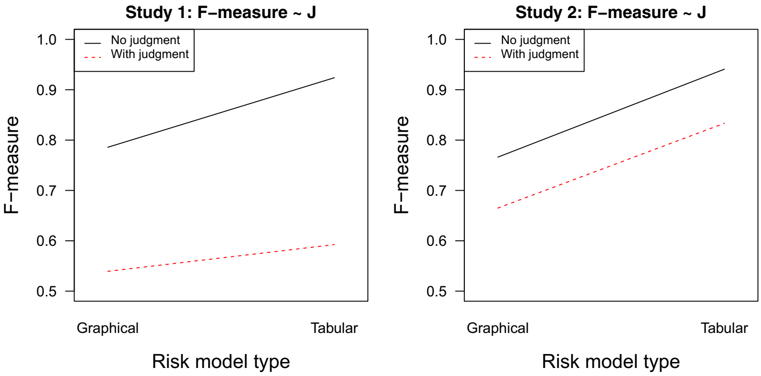


Fig. 11 Effect of complexity (J) on F-measure

References

Abrahao S, Gravino C, Insfran E, Scanniello G, Tortora G (2013) Assessing the effectiveness of sequence diagrams in the comprehension of functional requirements: Results from a family of five experiments 39(3):327–342

Agarwal R, De P, Sinha AP (1999) Comprehending object and process models: An empirical study 25(4):541–556

BSI (2012) Standard 100-1: Information Security Management Systems

De Gramatica M, Labunets K, Massacci F, Paci F, Tedeschi A (2015) The role of catalogues of threats and security controls in security risk assessment: An empirical study with ATM professionals. Springer

De Lucia A, Gravino C, Oliveto R, Tortora G (2010) An experimental comparison of ER and UML class diagrams for data modelling 15(5):455–492

Dunning D, Johnson K, Ehrlinger J, Kruger J (2003) Why people fail to recognize their own incompetence 12(3):83–87

Fabian B, Gürses S, Heisel M, Santen T, Schmidt H (2010) A comparison of security requirements engineering methods 15(1):7–40

Fox J, Weisberg S (2011) An R Companion to Applied Regression, 2nd edn. Sage, Thousand Oaks, CA. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Giorgini P, Massacci F, Mylopoulos J, Zannone N (2005) Modeling security requirements through ownership, permission and delegation. IEEE, p. 167–176

Grondahl IH, Lund MS (2011) Reducing the effort to comprehend risk models: Text labels are often preferred over graphical means 31:1813–1831

Hadar I, Reinhartz-Berger I, Kuflik T, Perini A, Ricca F, Susi A (2013) Comparing the comprehensibility of requirements models expressed in use case and tropos: Results from a family of experiments 55(10):1823–1843

Heijstek W, Kühne T, Chaudron MR (2011) Experimental analysis of textual and graphical representations for software architecture design. IEEE, p. 167–176

Hogganvik I, Stolen K (2005) On the comprehension of security risk scenarios. IEEE, p. 115–124

Hoisl B, Sobernig S, Strembeck M (2014) Comparing three notations for defining scenario-based model tests: A controlled experiment. IEEE, p. 95–104

Hothorn T, Hornik K (2015) exactRankTests: Exact Distributions for Rank and Permutation Tests. https://CRAN.R-project.org/package=exactRankTests, r package version 0.8-28

Kabacoff R (2015) R in action: data analysis and graphics with R. Manning Publications Co

Kaczmarek M, Bock A, Heß M (2015) On the explanatory capabilities of enterprise modeling approaches. Springer, p. 128–143

Labunets K, Massacci F, Paci F, Tran LMS (2013) An Experimental Comparison of Two Risk-Based Security Methods. IEEE, p. 163–172

Labunets K, Paci F, Massacci F, Ragosta M, Solhaug B (2014a) A First Empirical Evaluation Framework for Security Risk Assessment Methods in the ATM Domain. SESAR

Labunets K, Paci F, Massacci F, Ruprai R (2014b) An experiment on comparing textual vs. visual industrial methods for security risk assessment. IEEE, p. 28–35

Landoll DJ, Landoll D (2005) The security risk assessment handbook: A complete guide for performing security risk assessments. CRC Press

Lund MS, Solhaug B, Stølen K (2011) A guided tour of the CORAS method. In: Model-Driven Risk Analysis, Springer, pp 23–43

MacKenzie IS (2012) Human-computer interaction: An empirical research perspective. Newnes

Massacci F, Paci F (2012) How to select a security requirements method? a comparative study with students and practitioners. Springer, p. 89–104

Matulevičius R, Mayer N, Mouratidis H, Dubois E, Heymans P, Genon N (2008) Adapting secure tropos for security risk management in the early phases of information systems development. Springer, p. 541–555

Mayer N, Rifaut A, Dubois E (2005) Towards a risk-based security requirements engineering framework. vol 5

Mayer N, Heymans P, Matulevicius R (2007) Design of a modelling language for information system security risk management. pp 121–132

Mead NR, Allen JH, Barnum S, Ellison RJ, McGraw G (2004) Software Security Engineering: A Guide for Project Managers. Addison-Wesley Professional

Mellado D, Fernández-Medina E, Piattini M (2006) Applying a security requirements engineering process. Springer, p. 192–206

Moody D (2009) The ”Physics” of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering 35(6):756–779

Mouratidis H, Giorgini P (2007) Secure tropos: a security-oriented extension of the tropos methodology 17(02):285–309

Ottensooser A, Fekete A, Reijers HA, Mendling J, Menictas C (2012) Making sense of business process descriptions: An experimental comparison of graphical and textual notations 85(3):596–606

Purchase HC, Welland R, McGill M, Colpoys L (2004) Comprehension of diagram syntax: an empirical study of entity relationship notations 61(2):187–203

R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ricca F, Di Penta M, Torchiano M, Tonella P, Ceccato M (2007) The role of experience and ability in comprehension tasks supported by uml stereotypes. pp 375–384

Saleh F, El-Attar M (2015) A scientific evaluation of the misuse case diagrams visual syntax 66:73–96

Scanniello G, Gravino C, Genero M, Cruz-Lemus J, Tortora G (2014a) On the impact of uml analysis models on source-code comprehensibility and modifiability 23(2):13

Scanniello G, Staron M, Burden H, Heldal R (2014b) On the Effect of Using SysML Requirement Diagrams to Comprehend Requirements: Results from Two Controlled Experiments. pp 433–442

Scanniello G, Gravino C, Risi M, Tortora G, Dodero G (2015) Documenting design-pattern instances: A family of experiments on source-code comprehensibility 24(3):14

Sharafi Z, Marchetto A, Susi A, Antoniol G, Guéhéneuc YG (2013) An empirical study on the efficiency of graphical vs. textual representations in requirements comprehension. IEEE, p. 33–42

Stoneburner G, Goguen A, Feringa A (2002) NIST SP 800-30: Risk management guide for information technology systems. http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf

Stålhane T, Sindre G (2008) Safety hazard identification by misuse cases: Experimental comparison of text and diagrams. pp 721–735

Stålhane T, Sindre G (2012) Identifying safety hazards: An experimental comparison of system diagrams and textual use cases. pp 378–392

Stålhane T, Sindre G (2014) An experimental comparison of system diagrams and textual use cases for the identification of safety hazards 5(1):1–24

Stålhane T, Sindre G, Bousquet L (2010) Comparing safety analysis based on sequence diagrams and textual use cases. pp 165–179

Svahnberg M, Aurum A, Wohlin C (2008) Using students as subjects – an empirical evaluation. IEEE, p. 288–290

Vessey I (1991) Cognitive fit: A theory-based analysis of the graphs versus tables literature 22(2):219–240

Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. http://ggplot2.org

Wickham H (2016) gtable: Arrange ’Grobs’ in Tables. https://CRAN.R-project.org/package=gtable, r package version 0.2.0

Wood RE (1986) Task complexity: Definition of the construct 37(1):60–82

Katsiaryna Labunets is a postodoctoral research fellow in the Department of Information Engineering and Computer Science, University of Trento, Italy. She received a Ph.D. in Information and Communication Technology from the University of Trento in April 2016. Her current research interests include empirical methods, cyber security, risk assessment, and comprehensibility of risk modeling notations.

Fabio Massacci is a full professor at the University of Trento. He has a Ph.D. in Computing from the University of Rome La Sapienza in 1998. He has been in Cambridge (UK), Toulouse (FR) and Siena (IT). Since 2001 he is in Trento. He has published more than 200 articles on security and his current research interests are in empirical methods for security, risk assessment and security economics. He was the European Coordinator of the multi-disciplinary research project SECONOMICS on socio-economic aspects of security. Contact him at Fabio.Massacci@unitn.it

Federica Paci is a lecturer in Cyber Security at Electronics and Computer Science Department of University of Southampton. She received a PhD in Information Technology from the University of Milan in February 2008. Her recent research interests are in cloud computing and empirical security. She published more than 60 contributions as papers in international conferences and journals, and chapters in international books.

Sabrina Marczak is an assistant professor in the Computer Science School at the PUCRS University, Brazil, where she is a researcher of the Munddos Research Group. She holds a PhD degree in Computer Science from the University of Victoria, Canada. Her research interests are in Requirement Engineering, Collaborative Software Development, and Empirical Software Engineering.

Flávio Moreira de Oliveira received his Doctor in Computer Science degree at UFRGS (1994). He is a senior lecturer at Faculty of Informatics at the Catholic University of Rio Grande do Sul, Brazil (PUCRS), since 1992. At PUCRS, he worked as head of the Computer Science undergraduate program, the Information Systems undergraduate program, and of the Software Engineering Agency. His current research interests are Software Testing and Verification and Empirical Software Engineering.