Preliminary Experiments on the Relative Comprehensibility of Tabular and Graphical Risk Models

Katsiaryna Labunets[a], Yan Li[b], Fabio Massacci[a], Federica Paci[c],
Martina Ragosta[d], Bjørnar Solhaug[b], Ketil Stølen[b], Alessandra Tedeschi[d]

[a]University of Trento, Italy, {name.lastname}@unitn.it; [b]SINTEF ICT, Norway, {name.lastname}@sintef.no;
[c]University of Southampton, UK, F.M.Paci@soton.ac.uk; [d]Deep Blue, Italy, {name.lastname}@dblue.it

***Abstract*—The ATM SESAR projects have invested a significant effort to define, besides tabular representations, graphical modeling notations to capture ATM architectural elements. A key question is whether this is worth the effort for security risk assessment. It is important to understand which representation provides better comprehension of threats, vulnerabilities, security countermeasures, as well as the relationships between them. In this paper we present a preliminary study on the comprehensibility of two risk modeling notations, involving students from Trento and Oslo universities. In particular, we assessed the effect of using graphical or tabular modeling notation on the actual comprehension of security risk models. The subjects were asked to answer eight comprehension questions about the risk assessment concepts (like threats, vulnerabilities or controls) represented using graphical or tabular notation. The results of the data analysis show no significant difference in actual comprehension. Further studies are required to strengthen the statistical significance and to investigate the extent to which the findings are relevant for more general contexts.**

Keywords - empirical study, security risk assessment, comprehension, cognitive fit

I. Introduction

Air Traffic Management (ATM) has undergone a big evolution during the recent years, and continues to do so. The Single European Sky (SES) legislative framework has completely modified operations introducing new operational concepts and regulatory constraints, while the Single European Sky ATM Research Program (SESAR) is introducing advanced technologies and innovative procedures to enhance safety and capacity of the future aviation in Europe [1]. Within this framework, security issues become more and more relevant for the whole ATM system. In this complex scenario, security must be understood in a broad sense, gate-to-gate, and in a comprehensive manner, addressing all types of threats and including all interested parties and stakeholders.

Security concerns are usually addressed by conducting security risk assessments at different stages of the system development lifecycle. There are several established international standards and guidelines that specify the process, goals and activities of risk assessment. However, there is no evidence basis for deciding which of these techniques is more effective for the assessment of complex ATM systems.

The EMFASE project aims to address this gap by providing an innovative framework to compare and evaluate in a qualitative and quantitative manner risk assessment methods for security in ATM.

In this paper we report on an experimental comparison of two different techniques for representing risk, namely tabular and graphical, focusing on comprehensibility. Tabular representations are currently the most used in the ATM domain, and they are adopted by the SESAR SecRAM methodology [2]. However, some ongoing work in SESAR projects is proposing new graphical models to support the system modeling from a security perspective and in the risk assessment process. Therefore, the results of the EMFASE comprehensibility experiment can provide useful insights to the design of an integrated tabular and graphical notation for security risk assessment in ATM.

The study consisted of an experiment carried out in two rounds during the fall of 2014. The first round was conducted with MSc students enrolled in the Security Engineering course at the University of Trento, while the second round was conducted with MSc students of the Model Engineering course held at the University of Oslo.

The paper is organized as follows. Section II introduces the risk modeling notations selected for our study. The research method is presented in Section III, while Section IV reports the results of the analysis. The threats to validity of our study are discussed in Section V. Section VI discusses related work, while we conclude in Section VII.

II. Background

Most security risk assessment methods follow the general process as defined by the ISO 31000 risk management standard [3]. This process involves establishing the analysis context, identifying, analyzing and evaluating risk, and finally the treatment of unacceptable risks. End-users need to decide which techniques to use in conducting the activities and documenting the results. A number of such techniques is listed and described in the IEC 31010 standard [4], and includes, for example, brainstorming, interviews, check-lists, event tree analysis, and cause-consequence analysis. In the experiment

presented in this paper we focused on two techniques for risk identification, specification and documentation, namely tables and graphical models. While the representation of information is different in tables and graphical models, they are similar in the sense of structuring the information in a precise way. Both formats are moreover based on a well-defined universe of discourse. The universe of discourse is the entities that are analyzed, reasoned about and documented, such as vulnerabilities, incidents, and treatments. More precisely, by *risk table* we mean the arrangement of security risk information in columns, where each column corresponds to an element of the universe of discourse and where each row relates a set of such elements to each other. Risk graphs are *visual graphs for risk modeling*, by which we mean a set of elements that precisely matches a well-defined universe of discourse. The elements are nodes and edges that are visualized by the use of graphical icons.

In order to conduct the experiment we selected one specific table notation and one specific graphical notation. The selected notations should be state-of-the-art or industry best practice, and of a maturity that meets the needs and requirements of an industrial setting. In order to enable comparison, the two notations should moreover be comparable regarding their semantics. To ensure this we selected two notations with similar expressiveness. The selected risk table is from the NIST SP 800-30 guide for conducting risk assessments [5] and the selected notation for graphical risk modeling is the CORAS language [6].

NIST SP 800-30 is a standard developed by the National Institute of Standards and Technology. Published as a special document formulated for information security risk assessment, it pertains especially to IT systems. The document is a recommendatory guideline for securing IT infrastructures from a purely technical perspective. It was one of the first risk assessment standards, and many other standards are influenced by it. It has been widely used for IT security risk assessment globally, and is relevant to any business with an IT component. The NIST guide actually comes with a number of tables, each supporting a specific task in the process.

CORAS, similarly, comes with a number of different kinds of diagrams. CORAS is a model-driven approach to risk assessment that is closely based on the ISO 31000 risk management standard. It consists of three tightly interwoven artefacts, namely the CORAS method, the CORAS language and the CORAS tool. The method follows a process of eight steps that complies with the risk assessment process of the ISO standard. The language is a graphical notation with various kinds of diagrams that are used throughout the process from beginning to end. While being a formal language with support rigorous analysis of the diagrams, the language was developed to facilitate communication between stakeholders involved in the assessment, including people with little technical background. The CORAS tool is basically a diagram editor for creating all kinds of CORAS diagrams.

For the experiment we used the NIST table template for adversarial and non-adversarial risk and we used the CORAS treatment diagrams, because these two give a summarized overview of the most important elements and results of a risk assessment. In order to ensure the same expressiveness of the two notations we needed to restrict the syntaxes to a minor extent, and we also inserted three columns to the NIST template. These columns are for impact, asset and security controls, all of which are already part of the required documentation using other tables from this NIST guide.

TABLE I: Mapping between CORAS and NIST terms

CORAS	NIST SP 800-30
Threat	Threat source
Vulnerability	Vulnerability
Threat scenario	Threat event
Unwanted incident	Impact
Asset	Asset
Likelihood	Overall likelihood
Consequence	Level of impact
Treatment	Security control

Table I gives an overview of the elements used in the two documentation formats. The terms in each row have the same meaning and were used to model the exact same elements in the experiment.

Figure 1a shows an example of the instantiation of a CORAS diagram where we have named each of the elements used in the experiment. Figure 1b shows the entries of the corresponding row in the NIST table in the given order.

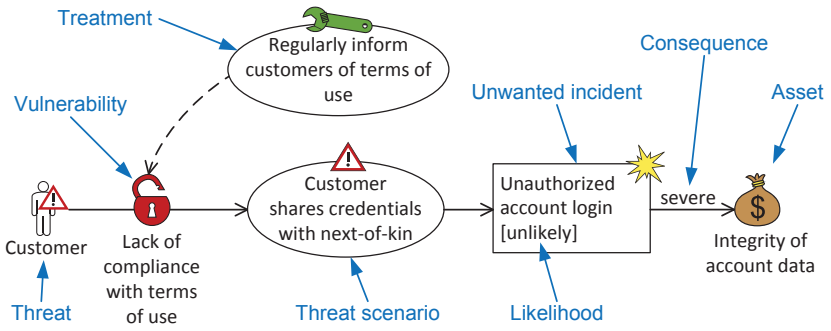
III. Research Method

We conducted the experiment by following established methods for empirical research [7], [8]. Our objective was explanatory, i.e. to seek an explanation in the form of a causal relationship [9] in order to investigate the effect of the risk model format on the comprehensibility of risk models. The experiment was done in two separate rounds, one with MSc students at the University of Trento as subjects, and one with MSc students at the University of Oslo. The dependent variable of the experiment was the comprehension of the risk models, whereas the independent variable that served as the treatment was the representation type of the risk model. All other independent variables had fixed levels in the experiment.

In order to study the comprehensibility of risk models as expressed in the two different formats, we investigated the following research question (RQ): *"Which risk model, the graphical one or the tabular one, is easiest to understand for subjects?"*

The comprehension level can be measured by precision and recall for each answer in the comprehension questionnaire. Moreover, a set of open questions allowed us to evaluate the answers using information retrieval metrics similar to De Lucia et al. [10]. By these metrics $answer_{s,i}$ is the set of answers given by participant s to question i, and $correct_i$ is the set of correct answers to question i.

$$recall_{s,i} = \frac{| \, answer_{s,i} \cap correct_i \, |}{| \, correct_i \, |}$$



(a) CORAS diagram

Threat event	Threat source	Vulnerability	Impact	Overall likelihood	Level of impact	Asset	Security control
Customer shares credentials with next-of-kin	Customer	Lack of compliance with terms of use	Unauthorized account login	Unlikely	Severe	Integrity of account data	Regularly inform customers of terms of use

(b) NIST table row entries

Fig. 1: Example of tabular and graphical models

$$precision_{s,i} = \frac{\mid answer_{s,i} \cap correct_i \mid}{\mid answer_{s,i} \mid}$$

To evaluate the comprehension level we used a measure that aggregates both precision and recall, namely the F-measure [11].

$$F\text{-}measure_{s,i} = 2 \cdot \frac{precision_{s,i} \cdot recall_{s,i}}{precision_{s,i} + recall_{s,i}}$$

We used the mean of the F-measures of all comprehension questions to evaluate the total comprehension level of the subject.

A. Comprehension Questions

The subjects were asked to fill a comprehension questionnaire regarding the contents of the risk models. Following analogous studies on comprehensibility of software engineering models [12], [13], [14], [10], the questionnaire consisted of eight questions aiming to test the ability of the subjects to identify a risk element of a specific type that is related to another element of a different type. Table II presents the exact comprehension questionnaire for graphical risk model that was used in the presented study. A corresponding, semantically equivalent, questionnaire was used for the tabular version.

B. Experiment Execution

The population of the controlled experiment was 35 students from the University of Trento and 11 students from the University of Oslo. After a five minutes introduction to the goals and objectives of the experiment, we gave the subjects a 10 minutes presentation to introduce the two kinds of risk notations, as well as the application scenario. The application scenario was an online banking scenario based on the Poste Italiane banking services that can be accessed via a web application or a smartphone app. The subjects answered the questionnaire from PCs using an online survey tool. The PCs were in the same room, but we made sure that subjects sitting next to each other were assigned different treatments.

TABLE II: Comprehension questions for graphical risk model

Q#	Question statement
1	Which threat scenarios can be initiated by exploiting vulnerability ”Weak malware protection”, according to the risk model? Please list all threat scenarios:
2	Which unwanted incidents are possible as a result of exploiting vulnerability ”Lack of mechanisms for authentication of app” by Cyber criminal? Please specify all unwanted incidents:
3	Which are the assets that can be harmed by the unwanted incident ”Unauthorized access to customer account via web application”? Please list all assets:
4	What is the likelihood that unwanted incident ”Unauthorized transaction via Poste App” occurs? Please specify the likelihood:
5	What is the lowest possible consequence for the asset ”User authenticity” that Cyber criminal can cause? Please specify the consequence:
6	Which threats can exploit the vulnerability ”Poor security awareness”? Please specify all threats:
7	What are the vulnerabilities that can be exploited to initiate the threat scenarios ”Cyber criminal alters transaction data” or ”Keylogger installed on computer”? Please list all vulnerabilities:
8	Which treatments are used to mitigate threat scenario ”Fake banking app offered on application store” or unwanted incident ”Unauthorized transaction via web application”? Please specify all treatments:

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

Procedure: The subjects were given five minutes to answer a demographics and background questionnaire, after which they had 20 minutes to complete the comprehensibility questionnaire. Half of the subjects were given NIST risk tables and half of them were given CORAS diagrams. The tables and CORAS diagrams represented the same security risk documentation (i.e. same semantics) as derived from the Poste Italiane application scenario. After the completion of the comprehensibility questionnaire, the subjects had two minutes to fill in the post-task questionnaire reported in Table III. The purpose of the post-task questionnaire was to control the possible effect of the experimental setting on the results.

TABLE III: 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.

The ethical considerations were handled by informing the subjects in advance about the purpose of the study and how the gathered data is used by whom. Anonymity and confidentiality of personal data is guaranteed, and the processing and storage of the data is used only for the purposes of the study.

Due to a technical problem with the settings of the online survey tool we had to discard responses of 24 subjects in the experiment in Trento. Thus, in the final analysis we used responses of only 11 subject from Trento. This problem was fixed for and not present in the part of the experiment conducted in Oslo.

Analysis: We validated the null hypothesis H_0 (about no difference in the actual comprehensibility between tabular and graphical risk models) using unpaired test because we had between-subject design when the subjects apply one of two treatments. We tested the normality of the data distribution using the Shapiro-Wilk test. We used unpaired t-test for normally distributed data, otherwise we used its non-parametric analog, the Mann-Whitney test. For all statistical tests we set the significance level $\alpha = 0.05$.

IV. Experiment Results

A. Risk Model Comprehensibility

Table IV reports the descriptive statistics for precision and recall, both for the individual comprehension questions and overall. Overall, the two risk models had similar precision, but the tabular risk model demonstrated slightly better recall than the graphical one. At the level of each comprehension question, the subjects who used tabular risk model showed better precision and recall, with some exceptions. For question Q8 the two risk models demonstrated equal precision and recall. For the precision of Q2, Q3 and Q7 the graphical risk model outperformed the tabular one. For questions Q3 and Q7 the graphical risk model had better recall that the tabular one.

Figure 2 presents the average precision and recall of the subjects' responses to the comprehension questions. Most of the subjects (54%) who used tabular risk model achieved higher precision than the median, while only 44% of the subjects who used graphical risk model had precision higher than the median value. With respect to the recall of the subjects' responses, 62% of the subjects who used tabular risk model had recall higher than the median value, and only 33% of the subjects who used graphical risk model achieved high recall.

TABLE IV: Precision and recall by questions and risk model

Q#	Tabular Mean	Tabular Med.	Tabular sd	Graphical Mean	Graphical Med.	Graphical sd
Precision						
Q1	1	1	0	0.78	1	0.44
Q2	0.83	1	0.33	0.89	1	0.33
Q3	0.83	1	0.3	0.93	1	0.15
Q4	1	1	0	0.78	1	0.44
Q5	0.85	1	0.38	0.78	1	0.44
Q6	1	1	0	0.89	1	0.33
Q7	0.72	0.67	0.32	1	1	0
Q8	1	1	0	1	1	0
Overall	0.9	0.92	0.08	0.88	0.88	0.13
Recall						
Q1	0.92	1	0.19	0.72	1	0.44
Q2	0.81	1	0.33	0.67	0.5	0.35
Q3	0.88	1	0.3	1	1	0
Q4	1	1	0	0.78	1	0.44
Q5	0.85	1	0.38	0.78	1	0.44
Q6	1	1	0	0.83	1	0.35
Q7	0.77	1	0.33	0.83	1	0.25
Q8	0.74	1	0.34	0.7	0.67	0.2
Overall	0.87	0.85	0.11	0.79	0.83	0.16

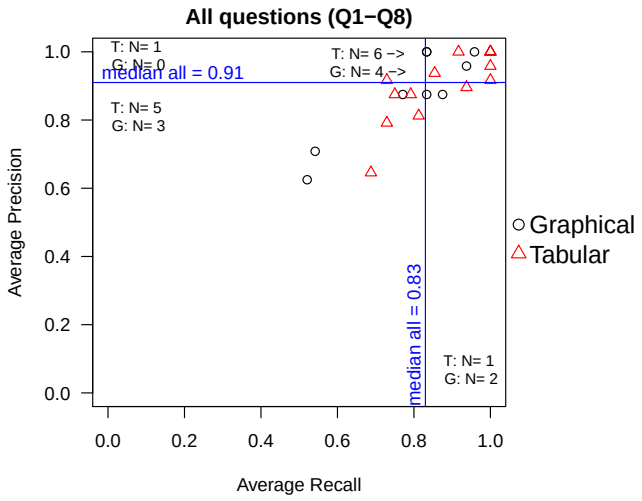


Fig. 2: Distribution of average precision and recall per subject by risk model type

Table V reports the mean, median, and standard deviation of precision, recall and F-measure by risk model type for each experiment round. To investigate the difference between two types of risk models we used Mann-Whitney test as our data is not normally distributed. The last two columns report Z statistics and p-value returned by the Mann-Whitney test.

The results of the test show that the subjects from Oslo demonstrated similar precision and recall using either graphical or tabular risk models. In contrast, the subjects from Trento showed better results using tabular risk model than using the graphical one. Overall, the subjects achieved equal level of actual comprehension of security risks (which we measured as

TABLE V: Average precision, recall and F-measure by experiments

	Tabular			Graphical			MW	
	Mean	Med.	sd	Mean	Med.	sd	Z	p-value
Precision								
SINTEF	0.9	0.93	0.09	0.94	0.96	0.06	0.83	0.40
UNITN	0.91	0.9	0.07	0.8	0.79	0.17	-1.16	0.25
Overall	0.9	0.92	0.08	0.88	0.88	0.13	-0.17	0.86
Recall								
SINTEF	0.85	0.83	0.09	0.87	0.88	0.08	0.55	0.58
UNITN	0.89	0.94	0.12	0.68	0.69	0.17	-1.53	0.13
Overall	0.87	0.85	0.11	0.79	0.83	0.16	-0.87	0.38
F-measure								
SINTEF	0.87	0.85	0.08	0.91	0.91	0.06	0.82	0.41
UNITN	0.9	0.92	0.1	0.74	0.73	0.17	-1.52	0.13
Overall	0.89	0.89	0.09	0.83	0.88	0.14	-0.6	0.55

TABLE VI: Post-task questionnaire results

	Tabular			Graphical		
Q#	Mean	Med.	sd	Mean	Med.	sd
Q1	0.46	0.00	0.88	0.22	0.00	0.44
Q2	1.38	1.00	1.04	1.00	1.00	0.87
Q3	0.77	1.00	0.73	0.67	0.00	0.87
Q4	0.62	1.00	0.51	0.67	1.00	0.71
Q5	0.54	0.00	0.66	0.78	1.00	0.83
Q6	0.54	0.00	0.66	0.78	1.00	0.83
Q7	0.38	0.00	0.51	0.67	1.00	0.71
Q8	0.23	0.00	0.44	0.44	0.00	0.53
Q9	Yes (62%) / No (38%)			Yes (22%) / No (78%)		

F-measure) using both tabular and graphical risk models. The Mann-Whitney test did not reveal any significant difference between the two risk models for all experiment and dependent variables. Thus, we cannot reject the null hypothesis H_0 for the scenario and circumstances studied in this experiment.

B. Post-task Questionnaire

We used the responses to the post-task questionnaire to control the possible effect of the experiment settings on the results. Table VI reports mean, median and standard deviation of the responses by risk model type. The responses are on a 5-item Likert scale from 0 (strongly agree) to 4 (strongly disagree). Overall, all subjects—regardless the type of risk model used—conclude that the settings were clear, the task was reasonable and the materials were clear and sufficient.

Note that most of the subjects (62%) who used tabular risk model reported that they used search in browser or MS Excel, while only 22% of the subjects who used graphical risk model reported that they used search in browser or PDF viewer.

C. Co-Factor Analysis

To test the possible effect of the co-factors on the dependent variables we used the two-way ANOVA, which is robust in case of violation of normality assumption, and is widely accepted in the literature for co-factor analysis [13], [15], [16]. We considered co-factors like work experience, level of expertise in security, modeling languages, as well as in the domain of online banking. Only one subject reported his knowledge in modeling languages as ”novice”. Therefore, we merged this category with the category ”beginner”. Another subject reported his knowledge of the online banking domain as ”proficient user”, and therefore, we merged this category with the ”competent user” category.

The results of two-way ANOVA revealed only one statistically significant interaction. There is an interaction between the experiment round (Trento and Oslo) and risk model type with the effect on the F-measure, and this is statistically significant according to the results of two-way ANOVA (p-value = 0.047). The F-measure results presented in Table V, comparing the findings from Trento and Oslo, clearly illustrate this effect. The two-way ANOVA also revealed a statistically significant effect of the subjects' level of expertise in modeling languages on the recall (p-value = 0.03). The results show that the subjects with higher level of expertise in modeling languages (e.g., ”proficient user”) provided more complete responses (median recall is 0.94) than the subjects with average or low level of expertise (median recall is 0.82).

V. Threats to Validity

Threats to validity can be structured into the four categories of internal validity, external validity, conclusion validity, and construct validity [17].

Threats to internal validity are factors that may affect the dependent variables and that have not been taken into account in the experiment. Because this experiment was conducted in two rounds, one at the University of Trento and one at the University of Oslo, there is a risk that the introduction to the experiment and the preparation of the participants differed to an extent that affected the results. Question Q9 (see Table III) furthermore revealed that 62% of the participants using the tabular used search, sort and copy-paste, while only 22% of the participants using graphical models did so. How this may affect the comprehensibility (for example by saving time) was not investigated. There is moreover a significant difference between the two experiment rounds in this respect, as only one participant (assigned to the tabular risk model) in Oslo used search and sort.

Threats to external validity regard the ability to generalize the results of the experiment. One issue here is whether the participants (MSc students) are representative for the target population (security risk assessors). It may, of course, be that the results could be different with participants with some experience from risk modeling. However, our experiment was designed to ensure that the two groups were given the exact same information and risk models with the same semantics, so that comprehensibility could be compared independent of any background of the participants. Another issue regarding external validity is the extent to which CORAS and NIST are representative for, respectively, graphical and tabular models in general. In particular, CORAS was designed to be intuitive and easy to understand and may therefore perform better than alternative graphical models. This is something that needs to be investigated in further experiments. There is also the question of whether the experiment results generalize to large-scale risk models. In real security risk assessments, the tables and graphical models may be significantly larger than the ones we used in the experiment. Comprehensibility of large

models is therefore another object for further studies. A further issue is that the models and comprehensibility questions that were used in the experiment are not representative for all kinds of uses of such risk documentation in security risk assessments. Consistency checking, identification of statistical dependencies, and likelihood estimation, for example, were not addressed in the study.

Conclusion validity regards the possibility to draw correct conclusions based on the results. An issue here is the statistical power. We plan to repeat the study and to conduct similar experiments to cope with this validity threat, and strengthen or revise the conclusions accordingly.

Construct validity regards the extent to which the studied measures actually represent what the researcher seeks to investigate. In this controlled experiment we carefully designed the study to ensure that threats to construct validity were eliminated, or at least insignificant regarding the experiment results.

VI. Related Work

There exists extensive work that compares textual and visual notations in software engineering. The existing works can be divided into a) works that propose cognitive theories to explain the differences between the two notations or that emphasize their relative strengths (e.g [18], [19]); b) works that compare the two notations from a conceptual point of view; and c) works that empirically compare visual and textual notations (e.g [20], [21]). In the following we discuss the latter only, as it concerns empirical studies on the effectiveness of visual vs. textual notions in supporting software engineering tasks.

Sharafi et al. [20] assessed the effect of using graphical vs. textual representations on subject's efficiency in performing requirements comprehension tasks. They found no difference in the accuracy of the answers given by subjects when using the textual and the graphical notation. However, the subjects preferred the graphical notation, even though it took them considerable more time to perform the task than when using the textual one.

Similarly, Stålhane et al. conducted a series of experiments to compare the effectiveness of textual and visual notations in identifying safety hazards during security requirements analysis. In [22], they 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 than use-case diagrams. In more recent experiments [21], [23], [24], Stålhane et al. compared textual misuse cases with UML system sequence diagrams. The experiments revealed that textual misuse cases are better than sequence diagrams when it comes to identifying threats related to required functionality or user behavior. In contrast, sequence diagrams outperform textual use cases when it comes to threats related to the system's internal working.

Heijstek et al. [25] 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 terms of communicating software architecture design decisions. Ottensooser et al. [26] compared understandability of textual and graphical notations for business process description. The authors chose written use cases and BPM notation as instances of textual and graphical notation, respectively. The results showed that all subjects well understood the written use cases, while the BPMN models were well understood only by the students with good knowledge of BPMN.

While there are empirical studies that compare graphical and textual representations for requirements [20], [22], [21], [23], [24], software architectures [25], and business processes [26], to the best of our knowledge, studies that focus on comparing textual and visual notations for security risk models are lacking. We have started to fill this gap by investigating the effectiveness and perception of textual and visual methods for security risk assessment in two previous empirical studies [27], [28]. We found that, in the setting of the particular studies, there is only a slight difference in the effectiveness of textual and visual methods, while the visual methods scored higher on preference than the textual methods. In this paper we have continued this work by exploring the relative comprehensibility of tabular and graphical risk models, which are representatives of textual and visual methods, respectively.

VII. Conclusions

In this paper we have reported on the results of an experimental comparison of tabular and graphical risk models regarding the comprehensibility of security risk documentation. The reported study is one of a series of experiments that we are conducting to investigate the suitability of security risk documentation both for reading the models and for reasoning about them. In designing the experiments we make use of real-world scenarios from the online banking domain and the ATM domain.

The presented study was an experiment with MSc students from the University of Oslo and the University of Trento. Overall, the results show that subjects achieve the same level of comprehensibility of security risks with the tabular and the graphical risk models. There are, however, differences at the level of individual experiments that indicate the need for additional work. Further experiments are required also to strengthen the statistical power and to study the extent to which the results are externally valid and can be generalized. It is, for example, and open question how the two approaches compare when the risk models are scaled up to match the size of the risk documentation of full security risk assessment. Other kinds of typical uses of risk assessment documentation also need to be investigated in order to strengthen the external validity of the results.

Acknowledgment

The work presented in this paper was partially supported by the SESAR JU WP-E via the EMFASE project (12-120610-C12).

REFERENCES

[1] *European ATM Master Plan. Edition 2*, SESAR Joint Undertaking, 2012.

[2] *SESAR ATM SecRAM implementation guidance material. Project deliverable 16.02.03-D03*, SESAR Joint Undertaking, 2013.

[3] ISO/IEC, "31000 – risk management – principles and guidelines," 2009.

[4] ——, "31010 – risk management – risk assessment techniques," 2009.

[5] NIST, "Guide for conducting risk assessment," *Special Publication 800-30*, 2012.

[6] M. S. Lund, B. Solhaug, and K. Stølen, *Model-Driven Risk Analysis – The CORAS Approach*. Springer, 2011.

[7] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Springer, 2012.

[8] R. K. Yin, *Case Study Research: Design and Methods*, 5th ed. Sage publications, 2014.

[9] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical software engineering*, vol. 14, no. 2, pp. 131–164, 2009.

[10] A. De Lucia, C. Gravino, R. Oliveto, and G. Tortora, "An experimental comparison of ER and UML class diagrams for data modelling," *Empirical Software Engineering*, vol. 15, no. 5, pp. 455–492, 2010.

[11] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[12] G. Scanniello, M. Staron, H. Burden, and R. Heldal, "On the effect of using SysML requirement diagrams to comprehend requirements: results from two controlled experiments," in *Proc. of EASE'14*. ACM, 2014, pp. 433–442.

[13] I. Hadar, I. Reinhartz-Berger, T. Kuflik, A. Perini, F. Ricca, and A. Susi, "Comparing the comprehensibility of requirements models expressed in Use Case and Tropos: Results from a family of experiments," *Information and Software Technology*, vol. 55, no. 10, pp. 1823–1843, 2013.

[14] H. C. Purchase, R. Welland, M. McGill, and L. Colpoys, "Comprehension of diagram syntax: An empirical study of entity relationship notations," *International Journal of Human-Computer Studies*, vol. 61, no. 2, pp. 187–203, 2004.

[15] M. Torchiano, F. Ricca, and P. Tonella, "Empirical comparison of graphical and annotation-based re-documentation approaches," *Software, IET*, vol. 4, no. 1, pp. 15–31, 2010.

[16] F. Ricca, M. Di Penta, M. Torchiano, P. Tonella, and M. Ceccato, "The role of experience and ability in comprehension tasks supported by UML stereotypes," in *Proc. of ICSE'07*, vol. 7, 2007, pp. 375–384.

[17] C. Wohlin, M. Höst, and K. Henningsson, "Empirical research methods in software engineering," in *Empirical Methods and Studies in Software Engineering*, ser. LNCS. Springer, 2003, vol. 2765, pp. 7–23.

[18] I. Vessey, "Cognitive fit: A theory-based analysis of the graphs versus tables literature," *Decision Sciences*, vol. 22, no. 2, pp. 219–240, 1991.

[19] D. Moody, "The "physics" of notations: Toward a scientific basis for constructing visual notations in software engineering," *IEEE Transaction Software Engineering*, vol. 35, no. 6, pp. 756–779, 2009.

[20] Z. Sharafi, A. Marchetto, A. Susi, G. Antoniol, and Y.-G. Guéhéneuc, "An empirical study on the efficiency of graphical vs. textual representations in requirements comprehension," in *Proc. of ICPC'13*. IEEE, 2013, pp. 33–42.

[21] T. Stålhane, G. Sindre, and L. Bousquet, "Comparing safety analysis based on sequence diagrams and textual use cases," in *Proc. of CAISE'10*, ser. LNCS, vol. 6051, 2010, pp. 165–179.

[22] T. Stålhane and G. Sindre, "Safety hazard identification by misuse cases: Experimental comparison of text and diagrams," in *Proc. of MODELS'08*, ser. LNCS, vol. 5301. Springer, 2008, pp. 721–735.

[23] ——, "Identifying safety hazards: An experimental comparison of system diagrams and textual use cases," in *Proc. of BPMDS'12*, ser. LNBIP, vol. 113. Springer, 2012, pp. 378–392.

[24] T. Stålhane and G. Sindre, "An experimental comparison of system diagrams and textual use cases for the identification of safety hazards," *International Journal of Information System Modeling and Design*, vol. 5, no. 1, pp. 1–24, 2014.

[25] W. Heijstek, T. Kühne, and M. R. Chaudron, "Experimental analysis of textual and graphical representations for software architecture design," in *Proc. of ESEM'11*. IEEE, 2011, pp. 167–176.

[26] A. Ottensooser, A. Fekete, H. A. Reijers, J. Mendling, and C. Menictas, "Making sense of business process descriptions: An experimental comparison of graphical and textual notations," *Journal of Systems and Software*, vol. 85, no. 3, pp. 596–606, 2012.

[27] K. Labunets, F. Massacci, F. Paci *et al.*, "An experimental comparison of two risk-based security methods," in *Proc. of ESEM'13*. IEEE, 2013, pp. 163–172.

[28] K. Labunets, F. Massacci, F. Paci, and R. Ruprai, "An experiment on comparing textual vs. visual industrial methods for security risk assessment," in *Proc. of EmpiRE Workshop at RE'14*, 2014, pp. 28–35.