No Search Allowed: What Risk Modeling Notation to Choose?

Katsiaryna Labunets
Delft University of Technology
Delft, The Netherlands
k.labunets@tudelft.nl

ABSTRACT

[Background] Industry relies on the use of tabular notations to document the risk assessment results, while academia encourages to use graphical notations. Previous studies revealed that tabular and graphical notations with textual labels provide better support for extracting correct information about security risks in comparison to iconic graphical notation. [Aim] In this study we examine how well tabular and graphical risk modeling notations support extraction and memorization of information about risks when models cannot be searched. [Method] We present results of two experiments with 60 MSc and 31 BSc students where we compared their performance in extraction and memorization of security risk models in tabular, UML-style and iconic graphical modeling notations. [Result] Once search is restricted, tabular notation demonstrates results similar to the iconic graphical notation in information extraction. In memorization task tabular and graphical notations showed equivalent results, but it is statistically significant only between two graphical notations. [Conclusion] Three notations provide similar support to decision-makers when they need to extract and remember correct information about security risks.

CCS CONCEPTS

• **Software and its engineering** → **Risk management**; • **Human-centered computing** → *Empirical studies in visualization*;

KEYWORDS

Cyber security risk assesssment; cyber risk modeling; comprehension; memorization; controlled experiment

ACM Reference Format:
Katsiaryna Labunets. 2018. No Search Allowed: What Risk Modeling Notation to Choose?. In *ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (ESEM '18), October 11–12, 2018, Oulu, Finland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3239235.3239247

1 INTRODUCTION

Security risk assessment (SRA) is a complex activity that plays an important role in software and systems engineering as it helps to identify relevant risks, prioritize them, and find adequate countermeasures to mitigate those problems. Further, SRA results must be clearly communicated with stakeholders to benefit from the findings, and lead to the implementation of proposed recommendations and necessary decisions, e.g., selecting a proper cyber insurance product. Professionals highlighted the importance of communication as one of the critical features for SRA methods [26, Table 2]. Specifically, in large corporations decision-making process involves stakeholders with different backgrounds and visions. Therefore, it is critical to communicate security risk information in a straightforward and objective way. For this purpose, industrial practitioners mostly rely on tabular notations, e.g., ISO 27005, NIST 800-30, or BSI IT-Grundschutz standards. Academia bets on graphical notations like i* [14] and CORAS languages [29], or recently proposed approach by Li et al. [28] for visualizing information security threats. There are some exceptions, for example, academia proposed SREP method [33] based on tables, while industry applies Microsoft STRIDE [18] approach that uses Data Flow Diagrams.

Previous studies with students and professionals showed that tabular notation supports better extraction of correct information about security risk over the iconic graphical notation [23, 25]. However, those finding might not give a full picture and have some limitations of construct validity: the comprehension task could potentially be biased in favor of tabular notations and did not reveal comprehensibility potential of graphical representation. Therefore, the goal of this study is to 1) compare tabular and graphical risk models in more equal settings and 2) advance in evaluation at different comprehensibility facets, namely information extraction and memorization. In extraction task we address validity concern by providing both tabular and graphical risk models in the form of images that does not allow participants to search (or filter tables) in models' artifacts. The memorization facet aims at mitigating a possible look-up nature of the comprehension task as participants have to fulfill the task without the model. It also tests how well the different types of models support memorization of information about security risks from decision-maker viewpoint.

The results reported in this paper address the following question: *"Which security risk model is more effective in extracting and memorizing correct information about security risks?"* To answer this question we conducted two controlled experiments with 60 MSc and 31 BSc students who were asked to complete similar comprehensibility tasks with and without having security risk models. From the results, in information extraction task we observed that participants with tabular notation obtain precision and recall similar to participants who used iconic graphical notation, even though it was not possible to search or sort tabular model. In memorization task, participants with tabular notation showed slightly lower comprehensibility in comparison to participants who used iconic graphical or UML notations, but the difference is insignificant.

ESEM '18, October 11–12, 2018, Oulu, Finland

ACM ISBN 978-1-4503-5823-1/18/10…$15.00
https://doi.org/10.1145/3239235.3239247

2 RELATED WORK

There are three main research streams in the literature comparing textual and visual notations. The first stream includes studies proposing cognitive theories to explain the difference between notations and their relative strengths [34, 48]. The second stream consists of studies comparing various notations from a conceptual viewpoint [21, 38]. The last one contains empirical studies comparing graphical and textual representations for, e.g., business processes [36], software architectures [17], safety and system requirements [9, 42, 44–46]. Recently, there were published a few empirical studies examining representations for security risks [15, 19, 25, 50] or comparing graphical and tabular methods for security risk assessment in full scale application [22, 24, 27, 30].

Empirical Research of Software Modeling Notation. Abrahao et al. [1] presented a family of controlled experiments with 112 participants with different levels of experience to investigate the effectiveness of dynamic modeling in requirements comprehension. The findings suggest that requirements specifications complemented with dynamic models (sequence diagrams) improve the understanding of software requirements in contrast to using only specification document. Scanniello et al. [39] reported a meta-analysis of a family of 12 controlled experiments with students and professionals to study the effect of UML analysis models on source-code comprehensibility. As treatments, they provided participants with source code with and without UML analysis models. The findings suggest that using UML models harms source code understanding and increases the time necessary to complete comprehension task. Sharafi et al. [42] compared three requirements modeling notations (Tropos diagrams, structured textual representation and the mix of two) regarding their effect on requirements comprehension. They did not observe any significant differences between models in participants' response precision, but they found that participants who used mixed representation used significantly less time to complete the task in comparison to the participants with only textual or graphical models. The authors explained that the latter finding could be due to the learning effect.

Empirical Research of Safety Modeling Notations. A research group of Stålhane et al. made a significant contribution to comprehensibility research in safety domain. They conducted a family of controlled experiments [44–46] to compare how useful textual and graphical notations for identification of safety hazards in security requirements analysis. The authors provided participants with textual use cases with system sequence diagrams [45, 46] and misuse case diagrams with textual misuse cases [44]. The results showed that textual representation assists users in focusing on the relevant areas. Also, textual alternative demonstrated better results in the identification of threats related to functionality and user behavior, while diagrams helped in understanding the system's internal working functionality and identifying related threats. Recently, de la Vara et al. [8] investigated the comprehension of safety compliance needs with textual and UML representations. The results revealed a small positive effect of using UML activity diagrams on the average effectiveness and efficiency of participants in understanding compliance needs, but the difference was not statistically significant.

Empirical Research of Security Risk Modeling Notations. In the past decade, empirical studies of security risk model comprehension got more contribution from different research groups. Matulevičius [31] reported an experiment with 28 graduate students in Computer Science to compare BPMN, Secure Tropos and misuse cases risk-oriented modeling notations w.r.t. their comprehensibility. The outcomes showed that BPMN based models were the most comprehensible out of three, while Secure Tropos and misuse case models were almost equal. A possible limitation of the study is that comprehension was measured by a simple 'look up' questions (e.g., "what is the security criterion?"). Managers who get SRA models must understand not only individual threat actors or vulnerabilities but also the relationships between them. We tried to address this aspect in the design of our comprehension questions (see Sec. 3 on p. 3).

Hogganvik and Stolen [19] compared the comprehensibility of UML and CORAS models in two controlled experiments with students. The results showed that the participants who used CORAS gave slightly more correct responses and spent less time to answer questions. A possible constraint of the study is that participants had ~5 min to answer 4-5 questions. We addressed this issue by allocating 40 minutes to answer 12 comprehension questions in total. The weakness of this work is the focus on diagram-based notations. In our work, we filled this gap by comparing UML-based and iconic CORAS representations with a tabular notation which is widely used in industrial security standards (e.g., NIST 800-30, ISO 27001, SESAR SecRAM, UK HMG IS1).

Yildiz and Böhme [50] recently conducted a controlled experiment with 85 participants to investigate the effects of risk visualization on managerial decision making in information security. This study showed that supplementing a textual description of security decision problem with graphical representation improves risk perception and participants' confidence in decisions, but does not contribute to the comprehension of the problem or security investment decision. In our prior study [25] we also found that participants achieved better or equal comprehension of described risk scenarios with tabular and UML-based notations.

There are few significant difference and some similarities between this study and our previous works [23, 25] which we summarize in Table 1. The main contribution of this work is studying how well tabular and graphical risk modeling notations support memorization of information about security risks. This development was suggested by the reviewer of our journal paper [23] with the goal to mitigate a possible bias in the comprehension task in favor of tabular notation.

3 EXPERIMENTAL DESIGN

We define the goal of our study according to the Goal Question Metric (GQM) template by Basili [4]: We *analyze* risk model comprehensibility *for the purpose* of assessing tabular and graphical modeling notations *with respect to* the extraction and memorization of correct information about security risks *from the viewpoint of* the decision-maker *in the context of* MSc and BSc students from the Delft University of Technology. We define the following research questions for our study:

RQ1 Which representation (tabular vs. graphical) improves participants effectiveness in *extracting* correct information about security risks?

Table 1: Comparison with previous works

Paper	Similar	Different
Labunets et al. [23] (EMSE'17)	We used application scenario and comprehension questionnaire similar to the second study reported in [23]. Both studies involved MSc students.	We introduced the memorization task and added UML-like risk modeling notation that combines textual labels with graphical representation.
Labunets et al. [25] (ESEM'17)	We used the same application scenario, risk modeling notations, and comprehension questionnaire.	In addition to information extraction, we introduced the memorization task. The presented experiments involved MSc students, while an earlier study was conducted with IT professionals.

RQ2 Which representation (tabular vs. graphical) improves participants effectiveness in *memorizing* correct information about security risks?

Experimental Task. In the extraction and memorization parts we asked our participants to answer a set of questions about information described in a risk model. Each set included six questions of different complexity levels. An example of the question: "*Which threat events can be initiated by Cyber criminal to impact the asset "Confidentiality of customer data"? Please select all unique threat events that meet the conditions (one or more element maybe correct).*" The sets were comparable between each other as included one question per combination of complexity factors along Wood's theory of task complexity [49] (i.e. information cues, relationships, and judgment acts) as adopted in practice by Labunets et al. [23]. The complexity factor was used to allow comparability of task between experimental parts and provide the diversity of questions regarding notation concepts to be understood. We did not have a goal to investigate the effect of task complexity factor as our experimental design provide a too small sample size for this purpose.

Table 2 presents two sets of comprehension questions that we provided to participants with graphical risk models. Questions for the tabular risk model are identical (except for the instantiation of the names of the elements to the textual risk modeling notation).

To test memorization performance and control participants' access to the artifacts, we had to provide participants with a picture of an assigned model and disabled the possibility to save images via the context menu of the browser. Also, we provided participants with multi-choice options for each question which consisted of a list of all unique elements present in the model. The list contained only elements' names (sorted alphabetically) but not their types (e.g., threat or vulnerability) as this could introduce additional bias by reducing the role of the model in task execution. The reason behind this step is to reduce possible mistakes due to manual typing of responses in the memorization part and make it comparable with the extraction part. Our participants were provided with images of risk models and could not copy-paste information like it was possible in Labunets et al. [23].

Research Hypotheses and Data Collection. From our GQM goal, we derived a set of null and alternative hypotheses (see Table 3). We did not formulate one-sided hypotheses like in Labunets et al. [23] as this study significantly different from the previous works.

The *independent variable* of our study is a risk modeling notation (tabular, UML, and CORAS). The *dependent variable* is comprehension level of participants that we evaluated based on participants' responses to a set of comprehension questions. As participants had to answer questions with one or more options, to quantify the comprehension level we could use information retrieval metrics, namely *precision*, *recall*, and their harmonic combination, the *F-measure*. Since our comprehension task included more than one question and we needed a single measure of participants' comprehension level, we aggregated all responses to calculate precision, recall, and F-measure at the level of the individual participant:

$$precision_{m,s,q} = \frac{|answer_{m,s,q} \cap correct_q|}{|answer_{m,s,q}|}, \quad (1)$$

$$recall_{m,s,q} = \frac{|answer_{m,s,q} \cap correct_q|}{|correct_q|}, \quad (2)$$

$$F_{m,s,q} = 2 * \frac{precision_{m,s,q} \times recall_{m,s,i}}{precision_{m,s,q} + recall_{m,s,q}}, \quad (3)$$

$$F_{m,s} = \text{mean}(\cup_{q \in \{1 \ldots N_{questions}\}} F_{m,s,q}\} \quad (4)$$

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.

Application Scenario. We kept the same scenario as in our prior works [23, 25] to have some comparability with our previous findings and mitigate possible threats to external validity. This scenario describes online banking services provided through a home banking portal, a mobile application, and prepaid cards. It was developed by our industrial partner, a large Italian corporation offering integrated services in finance and logistics. See Giacalone et al. [13] for more details on the company's internal SRA process.

Risk Modeling Notations. Our selection criteria are: 1) comparability and 2) representativeness of studied notations and 3) coverage of core concepts used by the most common international security standards (e.g., ISO/IEC 27000 or NIST 800-30). Thus, we selected CORAS [29] as the most comprehensive graphical notation. This notation provides adequate coverage of central SRA concepts like an asset, threat, vulnerability, risk, and security control [11, 32]. Other possible candidates were ISSRM [32], Secure Tropos [35], and *si** [14]. The special feature of CORAS is a treatment overview diagram summarizing the SRA results. It is equivalent to summary tables in NIST's or ISO's standards.

As a tabular notation, we chose the table template for adversarial and non-adversarial risk from NIST 800-30 standard [43]. To show all the relevant information, we consolidated in a single table also the impact, asset and security control concepts (usually present in separate NIST tables).

For the mixed representation we used a UML-like modeling notation which replaces iconic elements in CORAS diagram by textual labels with element types. The related work [25] suggests that the availability of textual labels can help participants to understand risk models better and was found to be more preferred [15] over the pure graphical representation.

Figures 3a–3c in the appendix provide examples of fragments from CORAS and UML treatment diagrams, and NIST tables related to the risk of an HCN scenario that we used in the previous study.

Design. This experiment has a *between-subject design* where each participant completed a comprehension task using one of three

Table 2: Comprehension Questions for Graphical Risk Models

This table presents two set of comprehension questions provided to participants of the study with a graphical risk models (i.e. CORAS and UML). Questions for tabular model were identical up to renaming of the elements. Note: C - complexity level, IC - # of information cues, R - # of relationships, A - # of judgment acts.

Set 1

Q	C=$IC+R+A$	Question statement
1	2= 1 + 1 + 0	What are the consequences that can be caused for the asset "Availability of service"? Please select the consequences that meet the conditions (one or more element maybe correct).
2	3= 2 + 1 + 0	Which assets can be impacted by Hacker or System failure? Please select all unique assets that meet the conditions (one or more elements maybe correct).
3	4= 2 + 2 + 0	Which treatments can be used to mitigate attack paths which exploit any of the vulnerabilities "Poor security awareness" or "Lack of mechanisms for authentication of app"? Please select all unique treatments for all attack paths caused by any of the specified vulnerabilities (one or more elements maybe correct).
4	3= 1 + 1 + 1	What is the lowest consequence that can be caused for the asset "User authenticity"? Please select the consequence that meets the conditions (one or more elements may be correct).
5	3= 1 + 1 + 1	Which unwanted incidents can be caused by Hacker with "likely" or higher likelihood? Please select all unwanted incidents that meet the conditions (one or more elements maybe correct).
6	4= 2 + 1 + 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 (one or more elements may be correct).

Set 2

Q	C=$IC+R+A$	Question statement
1	2= 1 + 1 + 0	Which vulnerabilities can lead to the unwanted incident "Unauthorized transaction via Poste App"? Please select all vulnerabilities that meet the conditions (one or more elements may be correct).
2	3= 2 + 1 + 0	Which unwanted incidents can be caused by Cyber criminal with "severe" consequence? Please select all unwanted incidents that meet the conditions (one or more elements may be correct).
3	4= 2 + 2 + 0	Which threat scenarios can be initiated by Cyber criminal to impact the asset "Confidentiality of customer data"? Please select all unique threat scenarios that meet the conditions (one or more elements may be correct).
4	3= 1 + 1 + 1	Which threats can cause an unwanted incident with "severe" or higher consequence? Please select all threats that meet the conditions (one or more elements may be correct).
5	4= 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 select the lowest likelihood of the unwanted incidents that can be initiated using any of the specified vulnerabilities (one or more elements may be correct).
6	5= 2 + 2 + 1	Which vulnerabilities can be exploited by Hacker to cause unwanted incidents with "likely" or higher likelihood? Please select all vulnerabilities that meet the conditions (one or more elements may be correct).

Table 3: Experimental Hypotheses

Hyp	Null Hypothesis	Alternative Hypothesis
$H1$	No difference between notations in the level of comprehension when answering comprehension questions with available risk model (extraction task).	There is a difference between notations in the level of comprehension when answering comprehension questions with available risk model (extraction task).
$H2$	No difference between notations in the level of comprehension when answering comprehension questions without having a risk model (memorization task).	There is a difference between notations in the level of comprehension when answering comprehension questions without having a risk model (memorization task).

Table 4: Experimental design

Each participant was assigned to one of three models and question sets order. They used a corresponding model type to complete the assigned comprehension and memorization tasks on the scenario.

Group	Treatment	Questionnaires		Scenario
		Extraction part	Memorization part	
Group 1	Tabular	Set 1	Set 2	Online Banking
Group 2	Tabular	Set 2	Set 1	Online Banking
Group 3	UML	Set 1	Set 2	Online Banking
Group 4	UML	Set 2	Set 1	Online Banking
Group 5	CORAS	Set 1	Set 2	Online Banking
Group 6	CORAS	Set 2	Set 1	Online Banking

treatments: a tabular, CORAS, or UML risk models. Table 4 summarizes the experimental design of our study. Participants were randomly distributed between the three types of treatments and two sets of questions and worked individually. We chose this design for two reasons: 1) to eliminate a possible learning effect between treatments and 2) control a possible effect of different sets of questions. We also limited the time a single participant could spend on the overall experiment by 20 minutes as we used level of comprehension as performance metric [5]. The participation was anonymous and volunteer without any reward. Participants could withdraw from the experiment any moment before experiment completion.

Experimental Protocol. We used a three-phase protocol [30]:

- *Training:* Participants answered a brief individual demographics and background questionnaire and during 5 min watched a video tutorial on the appointed modeling notation and Online Banking application scenario.
- *Application:* The participants had to complete two parts:
  - *Part 1* was an extraction task where the participants had to review the appointed risk model and answer six comprehension questions. The order of questions was the same for all participants due to limitations of survey platform. Participants had 20 minutes to complete the task after which they were automatically advanced to the next page. An image of corresponding risk models was built in on the top of the task page and protected from downloading or opening in another tab in the browser. The tutorial on notation and scenario was provided at the beginning of the task and can be downloaded[1]. After finishing the task, participants filled in a post-task questionnaire.
  - *Part 2* was a memorization task where participants first need to memorize the same model as in part 1. After 5 minutes they were automatically forwarded to the comprehension task, but they no longer had access to the risk model and had to answer another six questions similar to part 1. The rest of the task was the same.
- *Evaluation:* Researcher checked the responses and marked correct and wrong answers to each comprehension question based on the predefined list of correct responses.

Data Analysis Procedure. To validate our null hypotheses we could use ANOVA test as we compare three treatments. However, ANOVA test makes assumptions regarding normality distribution (checked by Shapiro–Wilk test) and homogeneity of variance (checked by Levene's test) of our samples. In our case samples do not meet these requirements, we use the Kruskal–Wallis (KW) test and a post-hoc Mann–Whitney (MW) test (corrected for multiple tests with Bonferroni method). We adopt 5% as a threshold of α (i.e. the probability of committing a Type-I error).

[1]The full replication guide is available at https://sites.google.com/view/compr2017.

In case we fail to observe a statistically significant difference between treatments we can test their equivalence with TOST which initially was proposed by Schuirmann for testing the equivalence of generic and branded drugs [41]. The problem of the equivalence test can be formulated as follows:

$$\begin{array}{llll} H_{01} : \mu_A < \mu_B - \delta & \text{or} & H_{02} : \mu_A > \mu_B + \delta \\ H_{a1} : \mu_A \geq \mu_B - \delta & \text{and} & H_{a2} : \mu_A \leq \mu_B + \delta, \end{array} \quad (5)$$

where μ_A and μ_B are means of methods A and B, and δ corresponds to the range within which we consider two methods to be equivalent. The p-value is then the maximum among p-values of the two tests. The underlying test for each of the two hypotheses can then be any difference tests (e.g., t-test, Wilcoxon, etc.) as appropriate.

The FDA [12] recommends to use $\delta = [80\%; 125\%]$. On our bounded scale a percentage range could warrant statistical equivalence too easily when the mean value is close to the upper bound. Thus, we conservatively adopted $\delta = \frac{1}{2}\sigma = \pm 0.12$ that has been empirically derived by Labunets et al. [25] from related studies.

To control the effect of co-factors (e.g., working experience or level of English) on the actual comprehension in the form of F-measure, we use permutation test for two-way ANOVA, which is a suitable approach in case of violation of ANOVA's assumptions [20] (e.g., data has an ordinal type). The post-task questionnaire is used to control for the effect of the experimental settings and the documentation materials.

4 STUDY EXECUTION

The initial implementation of experimental setup has been tested in a pilot with several PhD students and faculty members at the Delft University of Technology (TU Delft). The initial experiment took place on September 14, 2017, at one of the lectures of the Cyber Risk Management course taught by a colleague of the author to MSc students in TU Delft. The replication of the experiment with BSc students occurred on September 18, 2017, as a part of the lecture at the Security & Organisation course. We collected 60 complete responses in the first experiment (20 with CORAS model, 20 with UML, and 20 with Tabular) and 31 in the second experiment (11 with CORAS, 9 with UML, and 11 with Tabular).

Table 5 presents the demographic and background information about participants in both experiments. Overall, our participants reported basic knowledge of requirements engineering, graphical modeling languages and security, and limited knowledge of risk assessment. Regarding the application scenario, they had a basic knowledge of online banking domain.

5 RESULTS

We begin with the analysis of the different experimental factors like the differences between experiments and sets of questions. We report the factors separately with a statistically significant difference, while factors without statistically significant differences were aggregated and analyze together.

Experiment. The permutation test for two-way ANOVA did not reveal any statistically significant interaction of experiments with modeling notation (p=1) nor effect on F-measure of participants' responses (p=0.80). Therefore, we analyze data collected in two experiments together.

Table 5: Demographic Statistics

The participants were 60 MSc and 31 BSc students attending courses at the TU Delft. Participants reported a good knowledge of English and basic knowledge of the related areas of expertise.

Variable	Scale	Mean/ Median	Distribution
Age	Years	22.6	19.8% were 19-20 years old; 38.5% were 21-22 years old; 33% were 23-24 years old; 8.8% were 25-37 years old
Gender	Sex		74.7% male; 25.3% female
Level of English	A1-C2		1.1% Pre-Intermediate (A2); 11% Intermediate (B1); 24.2% Upper-Intermediate (B2); 37.4% Advanced (C1); 26.4% Proficient (C2) or Native
Work experience	—	1.9	36.3% had no experience; 23.1% had 1 year or less; 22% had 2-3 years; 18.7% had 4 years or more
Expertise in requirements engineering	1(Novice)-5(Expert)	2 (median)	45.1% novices; 26.4% beginners; 24.2% competent users; 3.3% proficient users; 1.1% experts
Expert in graph. modeling languages	1–5	2 (median)	39.6% novices; 25.3% beginners; 25.3% competent users; 9.9% proficient users
Expert in risk assessment	1–5	1 (median)	64.8% novices; 25.3% beginners; 9.9% competent users;
Expertise in security	1–5	2 (median)	39.6% novices; 36.3% beginners; 19.8% competent users; 3.3% proficient users; 1.1% experts
Expert in online banking	1–5	2 (median)	42.9% novices; 37.4% beginners; 14.3% competent users; 4.4% proficient users; 1.1% experts

Table 6: RQ1 – Precision and recall by modeling notation

Tabular and CORAS models showed equivalent precision and recall. UML model showed similar precision as CORAS and tabular models, but had recall lower than the other two models.

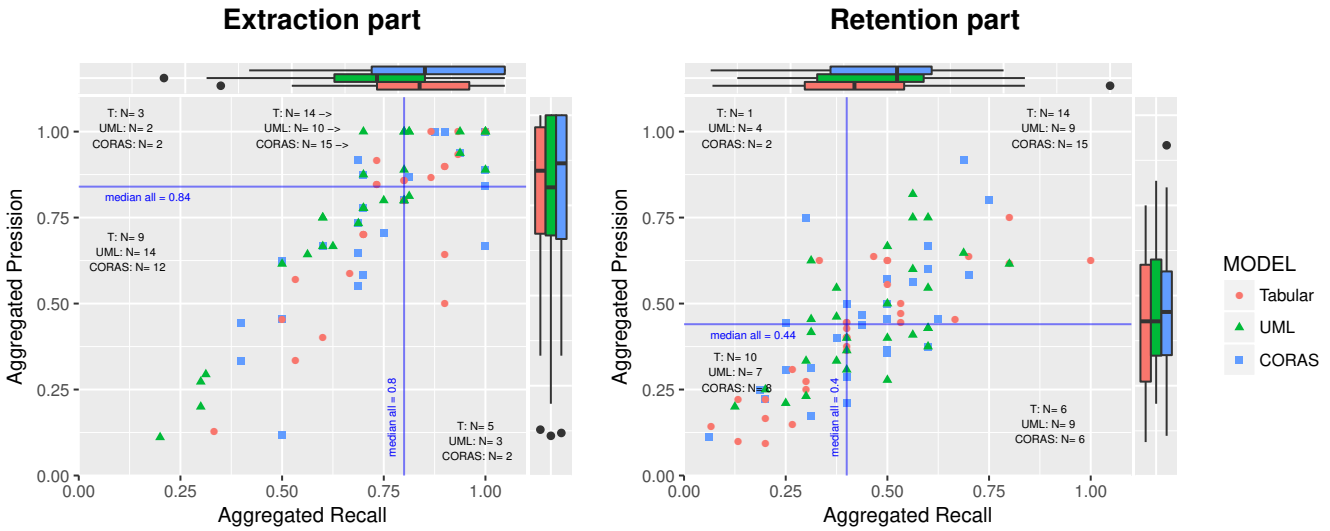
	Precision			Recall		
	mean	med	sd	mean	med	sd
Tabular	0.78	0.85	0.23	0.80	0.80	0.17
UML	0.76	0.80	0.25	0.69	0.70	0.21
CORAS	0.79	0.87	0.23	0.79	0.81	0.20

Experimental Task. We did not observe any statistically significant effect of two sets of questions (Set 1 and 2) on F-measure (permutation test for two-way ANOVA p= 0.10 in extraction and p= 0.60 in memorization parts) nor interaction with modeling notation (p=0.31 in extraction and p=0.79 in memorization parts). Thus, we analyze together the results of from two sets of questions. In this way, we eliminate a possible effect of a task order on the results.

Figure 1 compares the distribution of precision and recall of participants' responses in extraction (left) and memorization parts (right). If we take median precision and recall as a quality threshold for the level of comprehension, then we can see that 14/31 and 15/31 participants who used tabular and CORAS risk models respectively managed to reach the top right corner of the plot. In the case of UML, more participants appeared in the bottom left corner.

RQ1: Information Extraction. Table 6 presents descriptive statistics for precision and recall of responses to the extraction task. We can see the difference in favor of CORAS model over UML model (4% for precision and 16% for recall) and in favor of Tabular model over UML model (3% for precision and 14% for recall). The difference between CORAS and Tabular models is less than 1.5% both for precision and recall.

The results of the KW test did not reveal any statistically significant difference in precision and recall between three modeling notations (KW p> 0.17). We further investigated if there is a statistically significant equivalence between pairs of modeling notations using TOST with MW test with $\delta = \pm 0.12$. First, we tested the equivalence



Extraction part: Participants with tabular and CORAS risk models showed slightly better recall in contrast to the group who used UML risk model. However, the difference is not significant as evident from the overlapping boxplots on the top and right sides of the left figure. Regarding precision results, all models showed similar performance. *Memorization part:* There is a significant drop in precision and recall of participants' responses in comparison to the extraction part. The difference in precision and recall between three models is not significant as we can see from the overlapping boxplots on the sides of the right figure.

Figure 1: Participants' precision and recall by modeling notation and experimental part

of variance with Levene's test required by the MW test. The test confirmed that the samples have equal variances ($p > 0.18$). Table 7 sums up the findings of a post-hoc test with MW with Bonferroni correction ($\alpha = 0.05/3 = 0.017$) and TOST with MW test with correction suggested by Caffo et al. [6] ($\alpha = 0.05/(3^2/4) = 0.022$) for memorization task. For precision, we observed a statistically significant equivalence between three models. For recall, CORAS and Tabular models showed an equivalent result, and it was better than recall with the UML model. However, the difference between UML and other models in the recall is not statistically significant. Therefore, we can reject alternative hypothesis $H1_a$ for tabular vs. both graphical models for precision and tabular and CORAS models for recall. For other combinations, this question remains open.

RQ2: Information Memorization. Table 8 presents descriptive statistics for precision and recall of responses to the memorization task. We observed that both graphical modeling notation demonstrated 10-12% better response precision in memorization task than the tabular notation. The difference in the recall is smaller (7%) comparing to precision, but still in favor of graphical notations.

The results of the KW test did not reveal any statistically significant difference in precision and recall between three modeling notations (KW $p > 0.57$). Therefore, we investigated if there is a statistically significant equivalence between pairs of modeling notations using TOST with MW test. The Levene's test confirmed that the samples have equal variances (Levene's $p > 0.21$). Thus MW assumption holds for our samples. Table 9 summarizes findings of the statistical tests for memorization part. Regarding precision, we found that CORAS and UML models are equivalent with statistical significance, but for tabular and graphical models we got inconclusive results. In respect to recall, we obtained similar results. There is an equivalence between CORAS and UML models which was found statistically significant, while tabular and graphical models are equivalent at 10% significance level only because TOST test returned *p-value* = 0.029 and 0.041. Thus, we can reject an alternative hypothesis $H2_a$ only for the pair of graphical models, but not for pairs of tabular and any of two graphical models.

Post-task Questionnaire. We asked our participants to evaluate different aspects of study execution via the post-task questionnaire after each experimental part. Several questions were different between extraction and memorization parts. Table 10 presents descriptive statistics of participants' feedback. Responses are on a five-item Likert scale from 1 (strongly disagree) to 5 (strongly agree).

Overall, the participants found time to complete the task to be reasonable (question Q2) in both parts. The objectives of our study (Q3), task (Q4), and comprehension questions (Q5) were clear enough. Also, the participants did not struggle with understanding risk models (Q8) and using the electronic version of tabular and graphical models (Q9). They have a positive experience in using survey platform (Q10). Only participants who used tabular model reported 0.5 points lower responses regarding their experience with the Qualtrics platform. The difference is likely caused by the fact that tabular model was available in the form of a picture rather than a searchable document which is not a problem for graphical models. In extraction, part participants reported no significant difficulties in answering comprehension questions (Q6 in Table 10a), but in memorization part, the same questions were more challenging to the participants (Q6 in Table 10b). Also, participants were not sure if they had enough time to memorize risk model (Q1) and report about problems in model memorization (Q7). Higher cognitive load in memorization task comparing to extraction part can explain this.

Co-factor Analysis. We used the permutation test for two-way ANOVA to investigate the possible interaction between independent and dependent variables with several co-factors: participants' level of English, working experience, the level of participants' knowledge of security engineering, risk assessment, requirements

Table 7: RQ1 – Precision and recall by modeling notation

(a) Precision

Findings	Mean Precision		*p-value*	
	Model A	Model B	MW	$TOST_{MW}$
Tabular $\gtrsim$ UML	0.78	0.76	0.84	**0.015**
Tabular $\lesssim$ CORAS	0.78	0.79	0.75	**0.019**
UML $\lesssim$ CORAS	0.76	0.79	0.76	0.03

(b) Recall

Findings	Mean Recall		*p-value*	
	Model A	Model B	MW	$TOST_{MW}$
Tabular > UML	0.80	0.69	0.075	0.25
Tabular $\gtrsim$ CORAS	0.80	0.79	0.95	**0.006**
UML < CORAS	0.69	0.79	0.10	0.33

Table 8: RQ2 – Precision and recall by modeling notation

Three models showed similar precision and recall, but only for CORAS and UML models are close enough to be equivalence with statistical significance.

	Precision			Recall		
	mean	med	sd	mean	med	sd
Tabular	0.41	0.43	0.18	0.42	0.40	0.21
UML	0.46	0.43	0.17	0.45	0.50	0.16
CORAS	0.45	0.45	0.18	0.45	0.50	0.16

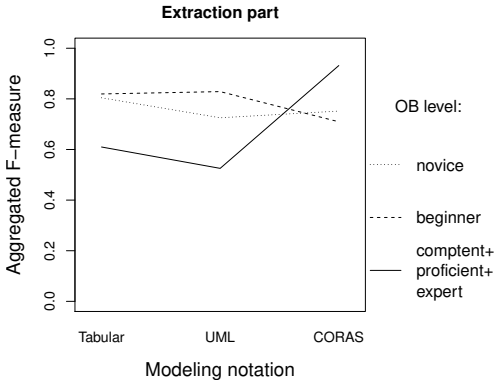


Figure 2: Interaction of treatments and domain expertise

engineering, graphical modeling languages, online banking. There was no statistically significant interaction between risk modeling type, dependent variables and any co-factor except one case. The test revealed a statistically significant interaction of participants' knowledge of online banking domain and modeling notation on F-measure in extraction part (p= 0.0046). We check this finding using the interaction plot presented in Figure 10a. As we had a small number of participants who reported their knowledge as "expert" (1 participant) and "proficient user" (4 participants) in online banking, we merged these categories with the category "competent user". We can see that the participants with lower levels of knowledge demonstrated a better overall level of understanding across all models. At the same time, participants with a higher level of knowledge showed worse comprehension level with tabular and UML models while the CORAS group demonstrated consistent results. A possible explanation could be the presence of Dunning-Kruger effect [10], when less competent people tend to overrate themselves higher due to a lack of competence and illusion about the level of their skills, while more competent people are likely to devalue their skills level as they think that others are more knowledgeable than themselves.

6 THREATS TO VALIDITY

Construct threats: Construct validity concerns whether the right metrics were used to investigate the comprehensibility of risk models. To mitigate this threat, we measured participants' level of comprehension using questionnaire and evaluated answers using information retrieval metrics (precision, recall, and F-measure) to avoid possible subjectivity in assessment. These metrics are widely adopted in the empirical software engineering literature [2, 16, 40]. The comprehension questionnaire was designed following a systematic approach inspired by related works[16, 37] and has been validated in our previous studies [23, 25]. Another relevant threat is the influence of the experimenter which we reduced by minimizing our involvement in the experimental process down to 10 minutes presentation about the high-level goals of the experiment and its procedure. The rest of the experiment was implemented using Qualtrics survey platform. Also, the decisions regarding experimental design were discussed with colleagues and tested in the pilot study to limit possible experimenter bias.

Internal Threats: Learning effects and order of task execution could threat internal validity. We mitigated it by adopting a within-subject design with a random assignment of subjects to the groups. Participants were instructed to fulfill task individually without interacting with other participants. To mitigate learning between extraction and memorization parts, we kept the same order of the parts for all participants. It was a feature of our experimental design to give participants enough time to get a hands-on experience with the model and learn it not only during 5 minutes before memorization task but also during the completion of the extraction part.

External Threats: Using students as experimental subjects could potentially harm the external validity, as they may be not representative enough for practitioners population. However, Svahnberg et al. [47] suggested that students can perform well in empirical studies. Moreover, we tried to recruit participants who have basic knowledge about security and modeling languages. To make the experimental settings as real as possible, our scenario was developed by an industrial financial company.

Conclusion Threats: A possible conclusion validity threat is related to the data analysis. We used the non-parametric tests as they do not require a normal distribution of the sample. To mitigate low statistical power, we adopted $\alpha = 0.05$ for the difference test and for the equivalence test $\delta = \pm 0.12$ that has been empirically determined by Labunets et al. [25] from related works.

7 DISCUSSION AND CONCLUSIONS

We summarize our findings as follows:

RQ1: *Which representation (tabular vs. graphical) improves participants effectiveness in extracting correct information about security risks?* The results revealed that the tabular model is equivalent to

Table 9: RQ2 – Summary of the findings

(a) Precision

Findings	Mean Precision		*p-value*	
	Model A	Model B	MW	$TOST_{MW}$
Tabular $\lesssim$ UML	0.41	0.46	0.44	0.07
Tabular $\lesssim$ CORAS	0.41	0.45	0.50	0.05
UML ~ CORAS	0.46	0.45	0.94	**0.01**

(b) Recall

Findings	Mean Recall		*p-value*	
	Model A	Model B	MW	$TOST_{MW}$
Tabular $\lesssim$ UML	0.42	0.45	0.34	0.041
Tabular $\lesssim$ CORAS	0.42	0.45	0.38	0.029
UML ~ CORAS	0.45	0.45	1.00	**0.004**

Table 10: Post-task questionnaire results

Extraction part: Participants agreed for all three model types that study objectives and task were clear, time was reasonable and provided models be clear enough.
Memorization part: Overall participants provided similar feedback on experimental settings except for questions related to model memorization and available time, and difficulty in replying comprehension questions.
Scale from 1 (strongly disagree) to 5 (strongly agree).

(a) Extraction part

	Tabular			UML			CORAS		
Q#	mean	med	sd	mean	med	sd	mean	med	sd
Q1	Not applicable in extraction part								
Q2	4.39	4.00	0.72	4.34	4.00	0.72	4.32	4.00	0.54
Q3	3.68	4.00	0.83	3.86	4.00	0.83	3.74	4.00	0.63
Q4	3.68	4.00	1.17	3.83	4.00	1.00	4.06	4.00	0.73
Q5	3.61	4.00	0.95	3.55	4.00	1.06	3.68	4.00	0.75
Q6	3.68	4.00	1.05	3.59	4.00	0.95	3.74	4.00	0.77
Q7	Not applicable in extraction part								
Q8	4.13	4.00	0.67	4.00	4.00	0.71	4.06	4.00	0.57
Q9	3.81	4.00	1.08	3.79	4.00	1.15	4.03	4.00	0.84
Q10	3.42	4.00	1.26	3.93	4.00	1.03	4.00	4.00	0.93

(b) Memorization part

	Tabular			UML			CORAS		
Q#	mean	med	sd	mean	med	sd	mean	med	sd
Q1	2.61	3.00	1.02	2.90	3.00	1.01	2.87	3.00	0.96
Q2	4.42	4.00	0.62	4.21	4.00	0.82	4.03	4.00	0.80
Q3	3.94	4.00	0.73	3.97	4.00	0.91	3.94	4.00	0.77
Q4	4.29	4.00	0.46	4.24	4.00	0.58	4.10	4.00	0.60
Q5	4.03	4.00	0.80	3.90	4.00	0.82	3.87	4.00	0.62
Q6	2.32	2.00	0.91	2.41	2.00	1.12	2.35	2.00	1.05
Q7	2.06	2.00	0.89	2.24	2.00	0.95	2.35	2.00	0.95
Q8	Not applicable in memorization part								
Q9	3.68	4.00	1.14	4.00	4.00	0.89	3.94	4.00	0.85
Q10	3.58	4.00	1.15	4.07	4.00	0.92	4.29	4.00	0.53

both graphical models with statistical significance for precision, but the equivalence between CORAS and UML is significant at 10% only. For recall, only tabular and CORAS models have statistically significant equivalence, while the other pairs showed some difference but not statistically significant. The UML notation showed lower recall compared to tabular (16% difference) and CORAS models (14%).

In contrast to our previous studies [23, 25], tabular representation did not show the best comprehension but performed at the level of other representations. The changes in the experimental settings could explain this. First of all, in our study the participants were not able to search provided models and filter tabular model which was available at our prior works. Previously this feature was extensively used by the participants with tabular (71% of participants) and CORAS models (70% of participants) (see responses to Q9 in post-task questionnaire in [25, Table IX]).

We also can notice that the level of precision (0.79 for CORAS group) in extraction task is similar to the precision of participants who used graphical models (overall precision 0.80-0.82) in two our previous studies with students (see precision results in [23, Tables 8-9]). In this study, our participants got a slightly better recall with CORAS (0.79) in comparison to the studies mentioned above where the participants who used CORAS model had an overall recall equals to 0.73 and 0.68, respectively in study 1 and 2. We suspect that multi-choice questions could have some contribution to better recall as it may provide a handy way to check that all relevant elements are selected in response to a specific question. This phenomenon requires additional research to be confirmed.

RQ2: *Which representation (tabular vs. graphical) improves participants effectiveness in memorizing correct information about security risks?* There is a small difference in the comprehensibility of three modeling notation in favor of graphical notations, but tests did not confirm statistically significant equivalence between tabular and graphical models in precision and simply at 10% significance level in response recall. Only two graphical models were found to be equivalent with statistical significance both in precision and recall. It can be explained that the UML and iconic risk models are equally good at supporting memorization of the correct information about security risks. Tabular notation provides a less clear representation of relations and presence of information duplication that affects participants' precision in responses.

Implications for research: This work contributes to the body of knowledge on model comprehensibility, specifically, for security risk management. We studied the effectiveness of tabular and graphical risk models in extraction and memorization of correct information about security risks. The results suggest that different type of comprehensibility task (extraction vs. retention) could expose different evaluation results [7].

Implications for practice: The main implication of our results for practitioners is the illustration of how well studies notation perform in different settings. If you need to present results in a fixed format (e.g., picture or slide), then both tables and diagrams could provide a similar level of information extraction and retain. However, if a decision-maker can work with risk model documents (e.g., search document) and does not need to remember all information, then tables are your best and more straightforward choice [25].

Future research: These days more and more information is delivered in electronic format. Thus, different scenarios of model communication and usage should be taken into account. For example, an experiment comparing comprehensibility and usability of a model snapshot vs. model file with available search and sort/filter function would fill this gap. Also, factors like level of participant's confidence in given responses and perceived difficulty of the task might shed some light on the comprehensibility of risk modeling notations as suggested by Aranda et al. [3].

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no 740920 (CYBECO). We would

like to thank B. Solhaug and K. Stølen from SINTEF for support in the definition of the CORAS and UML models, F. Paci from the University of Southampton and P. van Gelder, M. van Eeten, W. Pieters from the Delft University of Technology for their critique of experimental design and implementation.

REFERENCES

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

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

[3] Jorge Aranda, Neil Ernst, Jennifer Horkoff, and Steve Easterbrook. 2007. A framework for empirical evaluation of model comprehensibility. In *Proc. of MISE at ICSE 2007*. IEEE, 7–7.

[4] Victor R. Basili, Gianluigi Caldiera, and H. Dieter Rombach. 1994. The Goal Question Metric Approach. In *Encyclopedia of Software Engineering*, John J. Marciniak (Ed.). Vol. 1. John Wiley & Sons.

[5] Gunnar R Bergersen, Dag IK Sjøberg, and Tore Dybå. 2014. Construction and validation of an instrument for measuring programming skill. *IEEE Trans. Soft. Eng.* 40, 12 (2014), 1163–1184.

[6] Brian Caffo, Carolyn Lauzon, and Joachim Röhmel. 2013. Correction to "Easy Multiplicity Control in Equivalence Testing Using Two One-Sided Tests". *The American Statistician* 67, 2 (2013), 115–116.

[7] Nelly Condori-Fernández, Maya Daneva, Klaas Sikkel, and Andrea Herrmann. 2011. Practical relevance of experiments in comprehensibility of requirements specifications. In *Proc. of EMPIRE at RE 2011*. IEEE, 21–28.

[8] Jose Luis de la Vara, Beatriz Marín, Clara Ayora, and Giovanni Giachetti. 2017. An Experimental Evaluation of the Understanding of Safety Compliance Needs with Models. In *Proc. of ER 2017*. Springer, 239–247.

[9] Jose Luis de la Vara, Beatriz Marín, Giovanni Giachetti, and Clara Ayora. 2016. Do Models Improve the Understanding of Safety Compliance Needs?: Insights from a Pilot Experiment. In *Proc. of ESEM 2016*. ACM, 32.

[10] David Dunning, Kerri Johnson, Joyce Ehrlinger, and Justin Kruger. 2003. Why people fail to recognize their own incompetence. *Curr. Dir. Psychol. Sci.* 12, 3 (2003), 83–87.

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

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

[13] Matteo Giacalone, Federica Paci, Rocco Mammoliti, Rodolfo Perugino, Fabio Massacci, and Claudio Selli. 2014. Security triage: an industrial case study on the effectiveness of a lean methodology to identify security requirements. In *Proc. of ESEM 2014*. ACM, 24.

[14] Paolo Giorgini, Fabio Massacci, John Mylopoulos, and Nicola Zannone. 2005. Modeling security requirements through ownership, permission and delegation. In *Proc. of RE 2005*. IEEE, 167–176.

[15] Ida Hogganvik Grøndahl, Mass Soldal Lund, and Ketil Stølen. 2011. Reducing the effort to comprehend risk models: Text labels are often preferred over graphical means. *Risk Analysis* 31, 11 (2011), 1813–1831.

[16] Irit Hadar, Iris Reinhartz-Berger, Tsvi Kuflik, Anna Perini, Filippo Ricca, and Angelo Susi. 2013. Comparing the comprehensibility of requirements models expressed in Use Case and Tropos: Results from a family of experiments. *Inform. Soft. Tech.* 55, 10 (2013), 1823–1843.

[17] Werner Heijstek, Thomas Kühne, and Michel R.V. Chaudron. 2011. Experimental Analysis of Textual and Graphical Representations for Software Architecture Design. In *Proc. of ESEM 2011*. IEEE, 167–176.

[18] Shawn Hernan, Scott Lambert, Tomasz Ostwald, and Adam Shostack. 2006. Threat modeling-uncover security design flaws using the STRIDE approach. *MSDN Magazine-Louisville* (2006), 68–75.

[19] Ida Hogganvik and Ketil Stølen. 2005. On the comprehension of security risk scenarios. In *Proc. of IWPC 2005*. IEEE, 115–124.

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

[21] Monika Kaczmarek, Alexander Bock, and Michael Heß. 2015. On the Explanatory Capabilities of Enterprise Modeling Approaches. In *Proc. of EEWC 2015*. Springer, 128–143.

[22] Katsiaryna Labunets, Fabio Massacci, and Federica Paci. 2017. On the equivalence between graphical and tabular representations for security risk assessment. In *Proc. of REFSQ 2017*. Springer, 191–208.

[23] Katsiaryna Labunets, Fabio Massacci, Federica Paci, Sabrina Marczak, and Flávio Moreira de Oliveira. 2017. Model comprehension for security risk assessment: an empirical comparison of tabular vs. graphical representations. *Empir. Soft. Eng.* (2017), 1–40.

[24] Katsiaryna Labunets, Fabio Massacci, Federica Paci, and Le Minh Sang Tran. 2013. An Experimental Comparison of Two Risk-Based Security Methods. In *Proc. of ESEM 2013*. IEEE, 163–172.

[25] Katsiaryna Labunets, Fabio Massacci, and Alessandra Tedeschi. 2017. Graphical vs. Tabular Notations for Risk Models: On the Role of Textual Labels and Complexity. In *Proc. of ESEM 2017*.

[26] Katsiaryna Labunets, Federica Paci, Fabio Massacci, Martina Ragosta, and Bjørnar Solhaug. 2014. A First Empirical Evaluation Framework for Security Risk Assessment Methods in the ATM Domain. In *Proc. of SIDs 2014*. SESAR.

[27] Katsiaryna Labunets, Federica Paci, Fabio Massacci, and Raminder Ruprai. 2014. An experiment on comparing textual vs. visual industrial methods for security risk assessment. In *Proc. of EMPIRE at RE 2014*. IEEE, 28–35.

[28] Eric Li, Jeroen Barendse, Frederic Brodbeck, and Axel Tanner. 2016. From A to Z: developing a visual vocabulary for information security threat visualisation. In *Proc. of GraMSec 2016*. Springer, 102–118.

[29] Mass Soldal Lund, Bjørnar Solhaug, and Ketil Stølen. 2011. A Guided Tour of the CORAS Method. In *Model-Driven Risk Analysis*. Springer, 23–43.

[30] Fabio Massacci and Federica Paci. 2012. How to select a security requirements method? a comparative study with students and practitioners. In *Proc. of NordSec 2012*. Springer, 89–104.

[31] Raimundas Matulevičius. 2014. Model Comprehension and Stakeholder Appropriateness of Security Risk-Oriented Modelling Languages. In *Proc. of BPMDS 2014*. Springer, 332–347.

[32] Nicolas Mayer, André Rifaut, and Eric Dubois. 2005. Towards a risk-based security requirements engineering framework. In *Proc. of REFSQ 2005*, Vol. 5.

[33] Daniel Mellado, Eduardo Fernández-Medina, and Mario Piattini. 2006. Applying a security requirements engineering process. In *Proc. of ESORICS 2006*. Springer, 192–206.

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

[35] Haralambos Mouratidis and Paolo Giorgini. 2007. Secure Tropos: a security-oriented extension of the tropos methodology. *Int. J. Softw. Eng. Know. Eng.* 17, 02 (2007), 285–309.

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

[37] Filippo Ricca, Massimiliano Di Penta, Marco Torchiano, Paolo Tonella, and Mariano Ceccato. 2007. The role of experience and ability in comprehension tasks supported by UML stereotypes. In *Proc. of ICSE 2007*. 375–384.

[38] Faisal Saleh and Mohamed El-Attar. 2015. A scientific evaluation of the misuse case diagrams visual syntax. *Inform. Soft. Tech.* 66 (2015), 73–96.

[39] Giuseppe Scanniello, Carmine Gravino, Marcela Genero, José A Cruz-Lemus, Genoveffa Tortora, Michele Risi, and Gabriella Dodero. 2018. Do software models based on the UML aid in source-code comprehensibility? Aggregating evidence from 12 controlled experiments. *Empir. Soft. Eng.* (2018), 1–39.

[40] Giuseppe Scanniello, Carmine Gravino, Michele Risi, Genoveffa Tortora, and Gabriella Dodero. 2015. Documenting Design-Pattern Instances: A Family of Experiments on Source-Code Comprehensibility. *ACM Trans. Soft. Eng. Meth.* 24, 3 (2015), 14.

[41] D.L. Schuirmann. 1981. On hypothesis-testing to determine if the mean of a normal-distribution is contained in a known interval. *Biometrics* 37, 3 (1981), 617–617.

[42] Zohreh Sharafi, Alessandro Marchetto, Angelo Susi, Giuliano Antoniol, and Yann-Gaël Guéhéneuc. 2013. An empirical study on the efficiency of graphical vs. textual representations in requirements comprehension. In *Proc. of ICPC 2013*. IEEE, 33–42.

[43] Gary Stoneburner, Alice Goguen, and Alexis Feringa. 2002. NIST SP 800-30: Risk management guide for information technology systems. http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf.

[44] Tor Stålhane and Guttorm Sindre. 2008. Safety Hazard Identification by Misuse Cases: Experimental Comparison of Text and Diagrams. In *Proc. MODELS 2008*. 721–735.

[45] Tor Stålhane and Guttorm Sindre. 2014. An Experimental Comparison of System Diagrams and Textual Use Cases for the Identification of Safety Hazards. *Int. J. Inform. Sys. Model Design* 5, 1 (2014), 1–24.

[46] Tor Stålhane, Guttorm Sindre, and Lydie Bousquet. 2010. Comparing Safety Analysis Based on Sequence Diagrams and Textual Use Cases. In *Proc. CAISE 2010*. 165–179.

[47] Mikael Svahnberg, Aybüke Aurum, and Claes Wohlin. 2008. Using students as subjects-an empirical evaluation. In *Proc. of ESEM 2008*. ACM, 288–290.

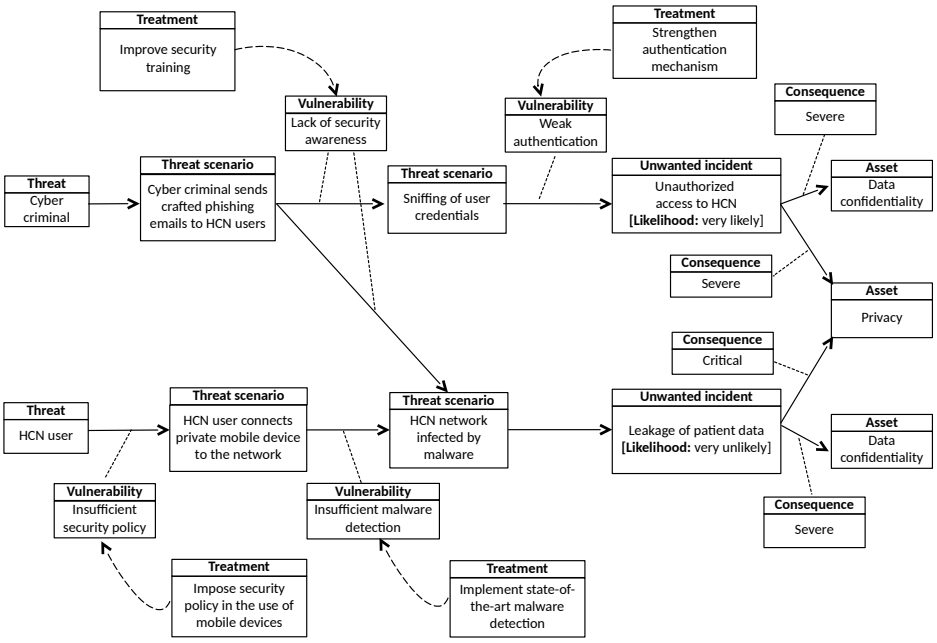
[48] Iris Vessey. 1991. Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature. *Decision. Sci.* 22, 2 (1991), 219–240.

[49] Robert E Wood. 1986. Task complexity: Definition of the construct. *Organ. Behav. Hum. Dec.* 37, 1 (1986), 60–82.

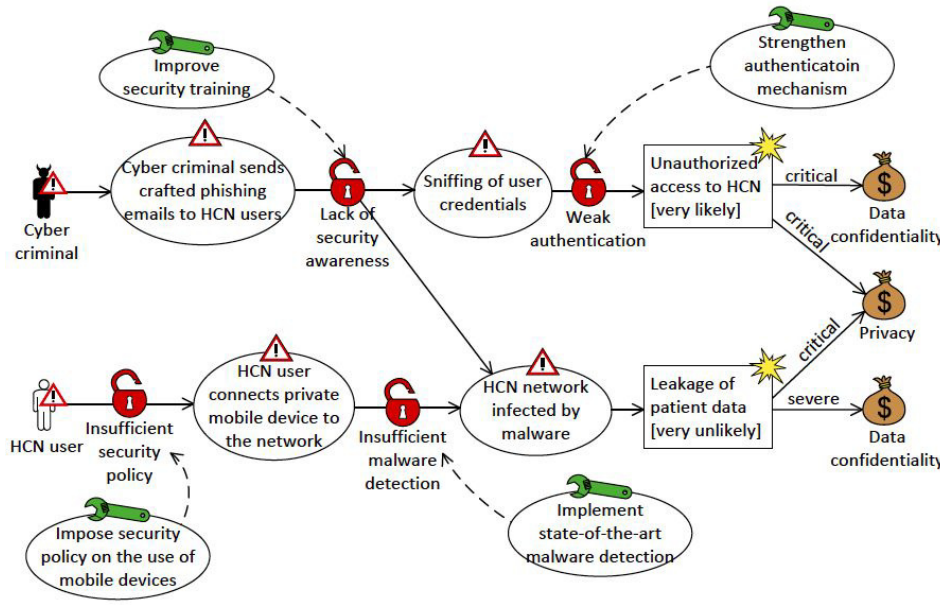
[50] Esra Yildiz and Rainer Böhme. 2017. Effects of information security risk visualization on managerial decision making. In *Proc. of EuroUSEC 2017*.

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.

(a) NIST table row entries



(b) UML diagram



(c) CORAS diagram

Figure 3: Fragment of a risk model in Tabular, UML-style, and CORAS notations