# Graphical vs. Tabular Notations for Risk Models: On the Role of Textual Labels

Labunets, Katsiaryna; Massacci, Fabio; Tedeschi, Alessandra

**DOI**
10.1109/ESEM.2017.40

**Publication date**
2017

**Document Version**
Accepted author manuscript

**Published in**
Proceedings of the 11th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2017

**Citation (APA)**
Labunets, K., Massacci, F., & Tedeschi, A. (2017). Graphical vs. Tabular Notations for Risk Models: On the Role of Textual Labels. In *Proceedings of the 11th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2017* (pp. 267-276). IEEE. https://doi.org/10.1109/ESEM.2017.40

# Graphical vs. Tabular Notations for Risk Models: On the Role of Textual Labels and Complexity

Katsiaryna Labunets[a,*], Fabio Massacci[b], Alessandra Tedeschi[c]

[a] *TPM, Delft University of Technology, The Netherlands*
[b] *DISI, University of Trento, Italy.*
[c] *Deep Blue srl, IT*

**Abstract**

[Background] Security risk assessment methods in industry mostly use a tabular notation to represent the assessment results whilst academic works advocate graphical methods. Experiments with MSc students showed that the tabular notation is better than an iconic graphical notation for the comprehension of security risks. [Aim] We investigate whether the availability of textual labels and terse UML-style notation could improve comprehensibility. [Method] We report the results of an online comprehensibility experiment involving 61 professionals with an average of 9 years of working experience, in which we compared the ability to comprehend security risk assessments represented in tabular, UML-style with textual labels, and iconic graphical modeling notations. [Results] Tabular notation are still the most comprehensible notion in both recall and precision. However, the presence of textual labels does improve the precision and recall of participants over iconic graphical models. [Conclusion] Tabular representation better supports extraction of correct information of both simple and complex comprehensibility questions about security risks than the graphical notation but textual labels help.

*Keywords:* Empirical Study, Security Risk Assessment, Risk Modeling, Comprehensibility, Cognitive Fit

## 1. Introduction

Vessey's paper (Vessey 1991) on how different cognitive tasks can be better achieved by different visual notations has sparked a long debate on what is a 'good' modeling notation. A field in which such debate is both active and relevant is Security Risk Assessment (SRA). Most academic approaches suggest a graphical notation, starting from Anti-Goals (Van Lamsweerde 2001) to the security extension of the i*-graphical requirements language (Giorgini et al. 2005) to highly iconic

*Corresponding Author. Accepted for publication in the proceedings of ESEM 2017 conference. The research leading to these results has been supported by the SESAR JU WPE under contract 12-120610-C12 (EMFASE). We would like to thank B. Solhaug and K. Stølen from SINTEF for support in the definition of the CORAS and UML models, F. Paci from University of Southampton for her help with the design of the study and questionnaire, and G. Frau and M. Ragosta from Deep Blue for their help in the implementation of the study and participants recruitment.

*Email addresses:* katsiaryna.labunets@unitn.it (Katsiaryna Labunets), fabio.massacci@unitn.it (Fabio Massacci), alessandra.tedeschi@dblue.it (Alessandra Tedeschi)

methods (Lund et al. 2011). Industry has adopted essentially tabular notations like OCTAVE, and the ISO 27005 and NIST 800-30 standards. Microsoft STRIDE (Hernan et al. 2006) is the industry exception whilst SREP (Mellado et al. 2006) is the academic one.

Such difference may be a mere time-lag (industry might eventually adopt graphical models) or actually caused by substantially different cognitive objectives (tabular notation works best for the prevalent cognitive tasks in industry).

In this respect, reported empirical studies have obtained conflicting results when using the method for *producing* an SRA. In short experiments, graphical methods fared better (e.g., (Hoganvik and Stølen 2005; Stålhane and Sindre 2008)). Studies with a full-fledged application for several days had either mixed results (e.g., (Massacci and Paci 2012; Labunets et al. 2013, 2014)) or concluded that the two notations were equivalent (Labunets et al. 2017).

Yet, producing a risk assessment may be the wrong cognitive task of interest for industry: security risk models are normally produced by (few) highly skilled experts but are consumed by (several) other actors, upwards to managers and downward to developers and operational staff. Hence, the *comprehensibility* of security risk models may be the key issue.

Previous experiments with MSc students (Labunets et al. 2017) have shown that the tabular notation from industry outperforms the graphical, highly iconic, notation from academia. Since there are several shades in graphical notations, we want to investigate whether a *mixed notation* combining textual labels with terse UML-style notation can achieve better results than either a purely tabular or an iconic graphical notation.

An important preliminary observation is that the success of a notation might also depends on the task's cognitive complexity as characterized by Wood's (Wood 1986), and adapted to the field by Labunets et al. (2017). According to Vessey's cognitive fit theory (Vessey 1991), by asking participants simple 'look up' questions we should favor tabular notations whilst asking them to identify 'spatially related' concepts might give graphical notation an advantage. Another issue to address is whether the answer could change when *professionals* would be asked to perform the task as opposed to students.

Hence in this paper we address the following questions for *participants with a significant work experience*:

*RQ*1 *Does the task complexity have an impact on the comprehensibility of the models?*

*RQ*2 *Does the availability of textual labels improve the participants effectiveness in extracting correct information about security risks?*

The short answer to *RQ*1 is that according to our experiment complexity has no impact and neither there is an interaction between notation and complexity. For *RQ*2, in terms of answer's precision the tabular notation is better but essentially equivalent to the UML-style with textual labels notation (roughly one wrong answer out of seven) and much better than the iconic graphical notation (approximately three wrong answers). In terms of recall, the tabular notation is definitely better than both graphical competitors.

## 2. Related Work

In the literature we found three main streams of works that compares textual and visual notations: *a)* studies that proposed cognitive theories to explain the differences between the notations or to explain their relative strengths (Vessey 1991), *b)* studies that compared different notations from a conceptual point of view (Kaczmarek et al. 2015; Saleh and El-Attar 2015), and *c)* studies

that empirically compare graphical and textual representations, e.g., for safety and system requirements (Sharafi et al. 2013; Stålhane and Sindre 2008; Stålhane et al. 2010; Stålhane and Sindre 2014; de la Vara et al. 2016), software architectures (Heijstek et al. 2011), and business processes (Ottensooser et al. 2012). To the best of our knowledge, there are few similar studies that empirically investigated modeling notations for security risk (Hogganvik and Stølen 2005; Grøndahl et al. 2011) or compared graphical and tabular security methods in full scale application experiments (Massacci and Paci 2012; Labunets et al. 2013, 2014, 2017).

### *2.1. Empirical Studies of Software Modeling Notations*

Abrahao et al. (2013) conducted a large scale study consisted of 5 controlled experiments with 112 participants with different levels of experience to evaluate the effectiveness of dynamic modeling in requirements comprehension. They found that requirements specifications supplemented by dynamic models (sequence diagrams) improves the comprehension of software requirements with respect to using only specification document. Scanniello et al. (2014) conducted four controlled experiments with students and professional to investigate the effect of UML analysis models on comprehensibility and modifiability of source-code. The treatments were providing participants with source code with and without UML analysis models. The results did not reveal any effect of using UML analysis models on understanding source code or ability to modify it. Similar to our paper, Sharafi et al. (2013) investigated three requirements modeling notations w.r.t. requirements comprehension. They compared Tropos diagrams, structured textual representation and mix of two. The results showed no differences between models in the accuracy of participants' responses, but they revealed that participants spent significantly less time to complete the task with the mixed model comparing to the textual and graphical models. The authors explained that the later finding could be due to the learning effect.

### *2.2. Empirical Studies of Security and Safety Modeling Notations*

Regarding the studies on comprehensibility in security domain, a series of controlled experiments were conducted by Stålhane et al. (Stålhane and Sindre 2008; Stålhane et al. 2010; Stålhane and Sindre 2014) to compare the effectiveness of textual and graphical notations in identifying safety hazards during security requirements analysis. They compared textual use cases with system sequence diagrams (Stålhane et al. 2010; Stålhane and Sindre 2014) and misuse case diagrams with textual misuse cases (Stålhane and Sindre 2008). These studies showed that the textual representation helps the user to focus in the relevant areas and was more effective in identification of threats related to functionalities and user behavior, whilst diagrams were more effective for understanding system's internal working and identifying related threats.

We found in the literature only three papers that empirically investigated the comprehensibility of security risk models. Matulevičius (2014) studied the comprehensibility of security risk-oriented modeling notations based on BPMN, Secure Tropos and misuse cases. These notations were extended with the concepts from the information security risk management (ISSRM) methodology (Mayer et al. 2005). The experiment involved 28 graduate students in Computer Science and showed that BPMN based models were the most comprehensible model over the other two, whilst Secure Tropos and misuse case models were almost equal. A limitation of the study is that comprehension has been measured by simple 'look up' questions asking to identify elements of a particular type in the model (e.g., "what is the security criterion?"). Managers who receive SRA models must understand not only individual threat actors or vulnerabilities, but also the relationships between

them. We tried to address this issue in the design of our comprehension questions (see Section 3 on p. 4).

Hogganvik and Stølen (2005) investigated the comprehensibility of UML and CORAS models in two controlled experiments with students: they found little difference in the correctness of participants' responses using CORAS over UML models and it took participants less time to answer questions with CORAS model than with UML model. However, the average time used to complete the task was around 5 min per set of questions, while in our study the participants had 40 minutes to answer 12 questions. Further, the study tested at once correctness and execution time and might be therefore suffer from construct validity. A more recent work (Grøndahl et al. 2011) studied the effect of textual labels and graphical icons (size, color, shape of elements) on the comprehension of risk models. The study with 57 IT professionals and students revealed that the participants preferred mixed models that combines textual and graphical representation over the pure graphical representation. Unfortunately, these two studies are focused only on diagram based notations. In our study we fixed this gap and compared UML-based and iconic CORAS notations with tabular representation as this is widely used in security industrial practice (e.g., NIST 800-30, ISO 27001, SESAR SecRAM, UK HMG IS1).

Our previous work (Labunets et al. 2017) studied the comprehensibility of tabular and graphical risk modeling notations and the effect of the task complexity on the level of comprehension. In comparison to this study we had several experiments, involved 152 MSc and BSc students as participants, and, as in previous studies, only compared tabular and iconic graphical notations. Tables better supported participants in extracting correct information about security risks than diagrams. This paper addresses a limitation of that study and other previous studies that might have played against iconic models: the presence of textual label marking elements (i.e. columns in tables) might favor the tabular representation when looking for relationship between elements. To validate whether such possibility is real, in this experiment we assigned a group of participants to use a UML-like model that provides diagrams with textual labels with elements' names. Our findings shows that such phenomenon is present.

## 3. Study Context and Planning

The goal of our study is to investigate the effect of task complexity ($RQ1$) and notation ($RQ2$) on the level of comprehension of information about security risks.

### *3.1. Comprehension task*

The comprehension task includes questions with different levels of complexity which varies along Wood's theory of task complexity (Wood 1986) as adapted to the field by Labunets et al. (2017).

The comprehension questions were designed taking into account the three main components of Task complexity:

- *Information cues (IC)* describe some characteristics that help to identify the desired element of the model. They are typically identified by a noun.
- *Relationships (R)* capture relations between a desired element and other elements of the model that must be identified to find the desired element.
- *Judgment acts (A)* require selecting a subset of elements meeting some criteria (e.g. "better").

Table 1: Comprehension Questions for Graphical Risk Models
This table presents the comprehension questionnaire provided to participants of the study with a graphical risk model. Questionnaires for Tabular and UML model where identical up to renaming of the elements. The full permutation of combinations is not possible because one relationship requires at least one information cue to bind one of the element of the relation (similarly for judgment acts).

| $Q$ | 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 specify the consequences that meet the conditions. |
| 2 | 2=1 + 1 + 0 | Which vulnerabilities can lead to the unwanted incident "Unauthorized transaction via App"? Please list all vulnerabilities that meet the conditions. |
| 3 | 3=2 + 1 + 0 | Which assets can be impacted by Hacker or System failure? Please list all unique assets that meet the conditions. |
| 4 | 3=2 + 1 + 0 | Which unwanted incidents can be initiated by Cyber criminal with consequence equal to "severe"? Please list all unwanted incidents that meet the conditions. |
| 5 | 4=2 + 2 + 0 | Which threat scenarios can be initiated by Cyber criminal to impact the asset "Confidentiality of customer data"? Please list all unique threat scenarios that meet the conditions. |
| 6 | 4=2 + 2 + 0 | Which treatments can be used to mitigate attack paths caused by any of the vulnerabilities "Poor security awareness" or "Lack of mechanisms for authentication of app"? Please list all unique treatments for all attack paths caused by any of the specified vulnerabilities. |
| 7 | 3=1 + 1 + 1 | What is the lowest consequence that can be caused for the asset "User authenticity"? Please specify the consequence that meet the conditions. |
| 8 | 3=1 + 1 + 1 | Which threats can impact assets with consequence equal to "severe" or higher? Please list all threats that meet the conditions. |
| 9 | 4=2 + 1 + 1 | Which unwanted incidents can be initiated by Hacker with likelihood equal to "likely" or higher? Please list all unwanted incidents that meet the conditions. |
| 10 | 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 specify the lowest likelihood of the unwanted incidents that can be initiated using any of the specified vulnerabilities. |
| 11 | 5=2 + 2 + 1 | Which vulnerabilities can be exploited by Hacker to initiate unwanted incidents with likelihood equal to "likely" or higher? Please list all vulnerabilities that meet the conditions. |
| 12 | 5=2 + 2 + 1 | What is the lowest consequence of the unwanted incidents that can be caused by Hacker and mitigated by treatment "Regularly inform customers of security best practices"? Please specify the lowest consequence that meets the conditions. |

We adopted Wood's formulation of the task complexity and calculate the *complexity of question* $i$ ($QC_i$) as follows:

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

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

As an example of computing task complexity, consider one of our comprehension questions: ($Q12$) "*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.*" The question complexity according to formula (1) is $2 + 2 + 1 = 5$ because there are two information cues ("Regularly inform customers of security best practices" for the element type "treatment", and "Hacker" for the element type "threat"), two relationships among them (A "consequence [...] caused by" B and C "mitigated by" D), and one judgment ("lowest consequence").

Table 1 presents the comprehension questionnaire that we provided to the participants of the study 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.

Table 2: Experimental Hypotheses

| Hyp | Null Hypothesis | Alternative Hypothesis |
|---|---|---|
| $H1$ | No difference between simple and complex questions in the level of comprehension (as measured by precision and recall of answers) when answering comprehension questions for all modeling notations. | The level of comprehensibility when answering simple comprehension questions is higher than for complex questions for all modeling notation. |
| $H2$ | No difference between notations with and without textual labels in the level of comprehension when answering comprehension questions. | The level of comprehension when answering comprehension questions using notations with textual labels is higher than using notations without textual labels. |

### *3.2. Research Hypothesis and Data Collection*

From our previous study is expected that the level of participants' comprehension of simple questions is higher than the comprehension of complex ones and that tabular notation performs better than the graphical notation. So, we can formulate a set of one sided alternative hypotheses (see Table 2).

The *independent variable* of our study is a risk modeling notation which can be either tabular, CORAS or UML. The *dependent variable* is the level of participants' comprehension which we assess based on participants' responses to a set of comprehension questions about information presented in risk models. Since we asked participants to respond questions with a subset of model's items, then $answer_{m,s,q}$ is the set of items provided in response to question $q$ by participant $s$ with modeling representation $m$ and $correct_q$ is the set of correct items which are expected for question $q$. Thus, to quantitatively evaluate the responses of our participants we can use information retrieval metrics, namely *precision*, *recall*, and their harmonic combination, the *F-measure*. These metrics are widely used in the empirical software engineering literature (Agarwal et al. 1999; Hadar et al. 2013; Scanniello et al. 2014, 2015). As our questionnaire contains more than one question and we would like to obtain a single value of participant's level of comprehension, we aggregate all responses to calculate precision, recall and F-measure at the level of the individual participant:

$$P_{m,s} = \frac{\sum_{q=1}^{N_{questions}} |answer_{m,s,q} \cap correct_q|}{\sum_{q=1}^{N_{questions}} |answer_{m,s,q}|}, \tag{2}$$

$$R_{m,s} = \frac{\sum_{q=1}^{N_{questions}} |answer_{m,s,q} \cap correct_q|}{\sum_{q=1}^{N_{questions}} |correct_q|}, \tag{3}$$

$$F_{m,s} = 2 * \frac{P_{m,s} \times R_{m,s}}{P_{m,s} + R_{m,s}}. \tag{4}$$

### *3.3. Application Scenarios*

We selected the same scenario from (Labunets et al. 2017) to allow the comparison with our previous study and to avoid introducing confounding factors: an online banking scenario developed by our industrial partner, a large Italian corporation offering integrated services in finance and logistics with a turnaround of around 24 billion Euro. The scenario describes the online banking services provided through a home banking portal, a mobile application and prepaid cards. See Giacalone et al. (2014) for a discussion of the company's internal SRA process.

### *3.4. Selection of Risk Modeling Notations*

There are several security risk modeling notations but to make this study fair and generalizable we need to find tabular and graphical notations that are i) comparable and ii) representative.

Table 3: Experimental design
Each participant was assigned to one of three groups and used a corresponding model type to complete the comprehension task on the scenario.

| Group | Treatment | Scenario |
|---|---|---|
| Group 1 | Tabular | Online Banking |
| Group 2 | UML | Online Banking |
| Group 3 | CORAS | Online Banking |

Selected notations should cover the core concepts used by the most common international security standards (e.g., ISO/IEC 27000 or NIST 800-30). In this regard, we chose CORAS (Lund et al. 2011) as the most comprehensive graphical notation. It provides a good coverage of the core SRA concepts, namely *asset, threat, vulnerability, risk, and security control* (Fabian et al. 2010; Mayer et al. 2005). We also considered graphical SRA methods like ISSRM (Mayer et al. 2005), Secure Tropos (Mouratidis and Giorgini 2007), *si∗* (Giorgini et al. 2005) and others. CORAS is the only method that has a one-diagram model summarizing the SRA results, the treatment overview diagram. It is the equivalent to the summary tables by NIST's or ISO's standards.

As a tabular notation we selected NIST 800-30 (Stoneburner et al. 2002) table template for adversarial and non-adversarial risk. NIST template gives an overview of the most important SRA elements. We consolidated in a single table also the impact, asset and security control concepts (present in separate NIST tables) to show all relevant information at once as in CORAS.

In this study we also introduced a UML-like modeling notation. This decision is motivated by *RQ*2 and aims to investigate the effect of textual labels on the actual comprehension of information presented in risk models. This experiment was actually suggested by the reviewers of our previous work (Labunets et al. 2017). Therefore, in collaboration with the designers of the CORAS language, we developed another version of the graphical risk model. This modeling notation is essentially the same as a CORAS diagram except that each iconic element of the CORAS language has been replaced by a uniform UML-style class element and an appropriate textual label.

Figures A.3a–A.3c in the appendix show fragments of CORAS and UML treatment diagrams, and NIST tables related to the risk of a HCN scenario that we used in previous study. Graphical models provide a global representation of several attacks by a "threat". Tabular model reports all possible attacks at the cost of duplicating information for the similar attacks which differ by one or two elements. A table only requires a simple navigation and provides look-up possibilities. The availability of textual labels with element types offer the same look/up benefits in comparison to having just graphical icons.

### *3.5. Experimental Design and Participants Recruitment*

This experiment has a *between-subject design* where each participant has been asked to complete comprehension task using one of three treatments: a tabular, CORAS, or UML model of security risks. Figure 3 summarizes the experimental design of our study. The participants were randomly distributed between the three type of treatments and worked individually. We chose this design type to eliminate possible learning effect between each treatment and because we could not control our participants as the experiment was conducted online. Also we wanted to limit the time a single participant would spend on the overall experiment.

After an initial pilot with some PhD students and post-docs at the University of Trento and the employees of Deep Blue, we sent email invitations across of network of contacts available to our research group at the University of Trento and Deep Blue. The email explained the high level goal

of the study, the task that the participants would be asked to do if they decide to participate, and the reward as a 20 Euro payment via PayPal. The link from the text message led to the web site with details about the study. Once participants clicked the "Start Experiment" button on the web site and consented to the experiment, they were randomly redirected by our script to one of three comprehension exercises implemented on SurveyGizmo, i.e. one task for each risk model type.

We used a three phases experimental protocol by Massacci and Paci (2012):

- *Training phase:* Participants answer short demographics and background questionnaire and see a video tutorial on the assigned modeling notation and the application scenario, namely an Online Banking scenario.
- *Application phase:* The participants are asked to review the risk model of the application scenario in the assigned representation and answer 12 comprehension questions. The order of the questions in the task is randomized for each participant. Participants are instructed to complete the task in 40 minutes. All necessary materials (e.g., risk model, tutorial slides) are provided at the beginning of the task. After the task completion, participants are asked to complete a simple post-task questionnaire.
- *Evaluation phase:* Researchers independently check the responses of the participants and code correct and wrong answers to each comprehension question based on the predefined list of correct responses.

### *3.6. Data Analysis*

To test the hypothesis $H1$ we can use one-sided paired t-test for normally distributed samples, or one-sided Wilcoxon test in case our samples are not normally distributed. As we have between-subject design with one factor and three treatments, we can use ANOVA test to validate the hypothesis $H2$. However, the ANOVA test makes assumption regarding normality distribution (checked by the Shapiro–Wilk test) and homogeneity of variance (checked by the Levene's test) of our samples. In case our samples do not meet these requirements we use the Kruskal–Wallis (KW) test and a post-hoc Mann–Whitney (MW) test (possibly corrected for multiple tests with Bonferroni method). We adopt 5% as a threshold of $\alpha$ (i.e. the probability of committing Type-I error).

In case we fail to observe statistically significant difference between treatments we can test their equivalence with TOST which initially was proposed by Schuirmann (1981) for testing the equivalence of generic and branded drugs. The problem of the equivalence test can be formulated as follows:

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

where $\mu_A$ and $\mu_B$ are means of methods $A$ and $B$, and $\delta$ 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 underlaying test for each of these two alternative hypothesis can then be any difference tests (e.g., t-test, Wilcoxon, etc.) as appropriate.

The Food and Drug Administration (2001) 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. Therefore, we define $\delta$ using an empirical approach. Four papers (Hadar et al. 2013; Scanniello et al. 2015, 2014; Abrahao et al. 2013) among the works discussed in Sec. 2 reported statistics for the F-measure and their aggregated variance is $\sigma_F = 0.23$ and we conservatively adopted $\delta = \frac{1}{2}\sigma = \pm 0.12$. The FDA range for the tabular notation would have been $[-0.18, +0.24]$ (see further Table 6).

Table 4: The number of participants reached each experimental phase
Each column is a subset of the previous columns. We discarded one participants from tabular group and one from CORAS group because they did not understand the task to be done.

| Treatment | Consented | Provided demographics | Finished task | Total valid |
|---|---|---|---|---|
| Tabular | 39 | 30 | 22 | 21 |
| UML | 30 | 23 | 20 | 20 |
| CORAS | 40 | 29 | 21 | 20 |
| Total | 109 | 82 | 63 | 61 |

To control the effect of co-factors (e.g., working experience or level of English) on the actual comprehension in 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 (Kabacoff 2015) (e.g., data has 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 experiment was conducted online in January-February 2016. All phases of experimental protocol were implemented on the SurveyGizmo platform. The participants were recruited through the mailing lists. In total, 572 participants accepted the consent form and moved to the actual comprehension task. Table 4 summarizes the number of participants in each treatment group that reached each experimental phase.

The completion rate is low (19%) but this is to be expected given that rewards were limited and subjects were professionals. Indeed, it has a similar rate to another security study where professionals were requested to code crypto API: 1571 people started the task, 660 dropped without taking any action, eventually only 337 (21%) arrived at the end (Acar et al. 2017).

Table 5 summarizes the demographic information about our participants that completed the task. Overall, they reported to have good general knowledge of architectural and system modeling and competent in the areas related to security, risk assessment, and graphical modeling languages. They also reported good competence in the application domain.

## 5. Addressing Threats to Validity

*Construct validity.* The design of our research instruments (comprehension questions, risk models) may affect the correctness of measuring the level of comprehension of information represented in different risk models. To mitigate this threat, these instruments were designed in collaboration with researchers from SINTEF who are the authors of CORAS notation and checked by five independent researches. The post-task questionnaire was adopted from the literature (Hadar et al. 2013; Ricca et al. 2007).

The design of comprehension questions may be a subject to bias in favor of tabular model as we used the names of elements in the questions. To control this threat, we introduced a representation that mixed tabular and CORAS notations, namely it provides UML-style diagrams with textual labels for elements' names (like tabular notation) instead of icons.

A critical problem for this experiment was encountered during its execution. Due to an implementation of the task on SurveyGizmo, the statements of five questions (Q4, Q6, Q8, Q9, and

Table 5: Demographic statistics
The participants were 61 professionals from 11 different countries with a good knowledge of English and a significant work experience.

| Variable | Scale | Mean/Med. | Distribution |
|---|---|---|---|
| Age | Years | 35 (mean) | 36% were 24–30 yrs old; 41% were 31–40 yrs old; 23% were 41–62 yrs old |
| Gender | Sex | | 74% male; 26% female |
| Education degree | BSc, MSc, MBA, PhD | | 11% have BSc degree; 36% MSc ; 8% MBA; 44% PhD degree |
| English level | A1–C2 | | 13% intermed. B1; 20% upper B2; 30% advanced C1; 33% proficient C2; 5% native |
| Work experience | Years | 9.6 (mean) | 3% did not report; 18% had 1–3 yrs; 43% had 4–7 yrs; 36% had >7 yrs |
| Expert in architectural and system specification and modeling | 0–4 | 2 (median) | 2% novices; 44% beginners; 23% competent users; 21% proficient users; 8% experts |
| Expert in sec. architecture and tech. | 0–4 | 3 (median) | 2% novices; 23% beginners; 25% competent users; 23% proficient; 23% expert |
| Expert in risk assessment | 0–4 | 3 (median) | 2% novices; 15% beginners; 30% competent users; 38% proficient; 13% expert |
| Expert in graph. modeling languages | 0–4 | 3 (median) | 2% novices; 21% beginners; 28% competent users; 26% proficient users; 15% experts |
| Expert in Online Banking | 0–4 | 3 (median) | 3% novices; 21% beginners; 21% competent users; 23% proficient users; 25% experts |

Q11) for graphical risk models were incorrect, namely the names of the concept elements were taken from the tabular notation. Many participants were able to successfully provide the correct responses to the questions as the names of concepts in tabular and graphical notations are very close. However, we decided to discard the responses to these question for all groups. All results and discussions reported in the paper are based on the responses to only seven unaffected questions (Q1-Q3, Q5-7, Q10, and Q12). This issue does not affect the overall results of the study, as we still have enough combinations, but barred a more refined analysis on which precise feature of task complexity (information cues, relationships, and judgment acts) that is most likely to impact the results.

*Conclusion validity.* We investigate possible effect of confounding factors on the results in order to assure that the difference in results is due to the treatment. Possible interaction between treatment and co-factors were tested by a permutation test for two-way ANOVA (Kabacoff 2015).

*Internal validity.* The simplicity of risk models might be threats to internal validity. The risk models for our study were designed by experts in CORAS and represent the realistic models reporting SRA results. The results of post-task questionnaire (see Table 9) indicate that participants well understood the objectives of the study and the tasks.

*External validity.* To make our findings generalizable we conducted the study with professionals with an average of 9 years of working experience and built our models based on realistic scenario provided to us by an industrial partner.

## 6. Results

### *6.1. RQ1: Task Complexity*

To test our research hypothesis $H1$ about the effect of task complexity, we aggregate F-measure by questions' complexity, e.g., $F_{m,s,\ell}$ is the precision value based on equation (4) for participant $s$

Table 6: F-measure by task complexity
The difference in F-measure between simple and complex questions is small across all participants ($N = 61$) for every notation. For the tabular notation complex and simple questions are statistically equivalent.

| | Simple | | | Complex | | | $W$ | $TOST_W$ |
|---|---|---|---|---|---|---|---|---|
| | mean | med | sd | mean | med | sd | $p$ | $p_{\delta=\pm 0.12}$ |
| Tabular | 0.94 | 1.00 | 0.15 | 0.91 | 1.00 | 0.16 | 0.30 | 0.001 |
| UML | 0.84 | 1.00 | 0.29 | 0.78 | 0.86 | 0.23 | 0.18 | 0.29 |
| CORAS | 0.66 | 0.67 | 0.31 | 0.63 | 0.69 | 0.31 | 0.28 | 0.10 |

Table 7: Precision and recall by modeling notation
Tabular model showed better precision and recall over both UML and CORAS notation. UML showed better precision and recall than CORAS.

| | Precision | | | Recall | | |
|---|---|---|---|---|---|---|
| | mean | med | sd | mean | med | sd |
| Tabular | 0.92 | 1.00 | 0.14 | 0.92 | 1.00 | 0.14 |
| UML | 0.83 | 0.91 | 0.17 | 0.77 | 0.83 | 0.21 |
| CORAS | 0.68 | 0.75 | 0.29 | 0.62 | 0.71 | 0.29 |

using risk model $m$ over all questions $q$ with complexity level $\ell$. Since $\ell = 2$ is the lowest possible level we aggregated questions as $\ell = 2$ and $\ell > 2$ (see complexity levels in Table 1).

Figure 1 compares the distribution of precision and recall for all questions (Figure 1a) and only complex ones (Figure 1b), i.e. the questions with complexity level greater than two. Table A.11 in the appendix also reports precision and recall of responses for each question. If we consider the median values of precision and recall as a quality threshold for the level of comprehension, then 13 out of 21 participants who used tabular model were able to reach the top right corner of comprehension plot. Most participants who used graphical model instead appears in the lower left corner.

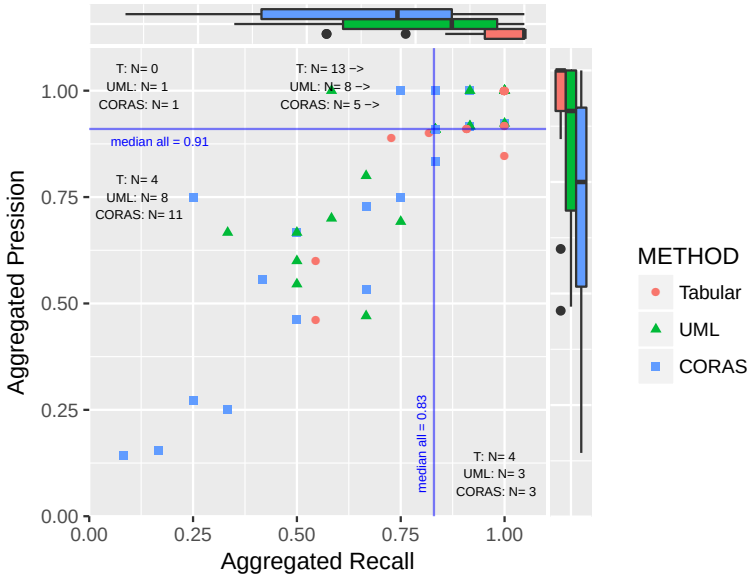
Table 6 presents the descriptive statistics for F-measure by questions' complexity. The results show that simple questions have better F-measure than complex ones. The difference in F-measure between two complexity levels varies between 3% for CORAS and tabular models and 6% for UML model.

The results of one-sided Wilcoxon tests did not reveal any statistically significant difference in F-measure between simple and complex questions for all three models. The results of the TOST with Wilcoxon test and $\delta = \pm 0.12$ revealed that for tabular model the level of comprehension of simple and complex questions is equivalent with statistical significance w.r.t. F-measure. Hence, we can reject the alternative hypothesis $H1_a$ for the tabular notation.

### *6.2. RQ2: Effect of notation on comprehension*

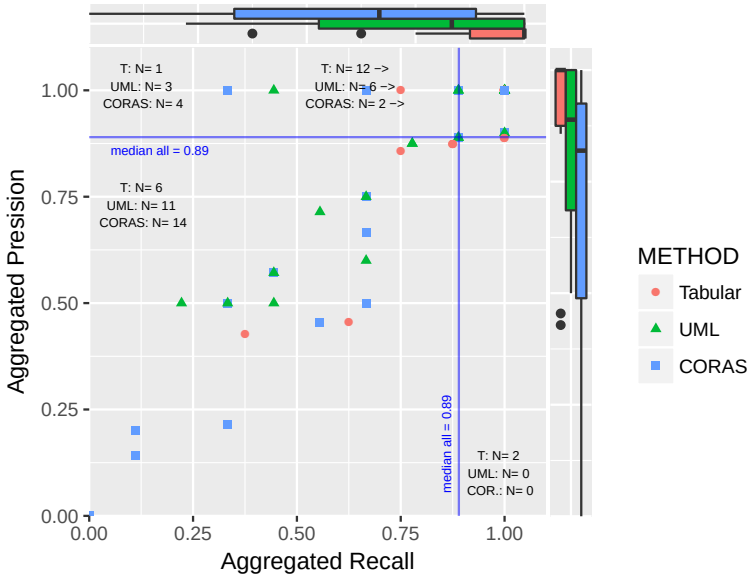
For RQ2 we report precision and recall separate as there is an important difference between modeling notations that can be flatten once aggregated in F-measure. Table 7 presents the descriptive statistics for precision and recall of responses to comprehension questions. Participants who used tabular model showed 9% better precision and 15% better recall than participants who used UML model and 24% better precision and 30% better recall than participants who used CORAS model. The difference between UML and CORAS results is 15% both for precision and recall in favor of participant who used UML.

The results of KW tests confirmed a statistically significant effect of risk model type both on precision (KW *p-value* = 0.009) and recall (KW *p-value* = 0.0002) of participants responses. As




(a) All questions

Participants with tabular risk model demonstrated better precision and recall in comparison to the participants who used CORAS risk model (see the non overlapping boxplots on the top and right parts of the figure). The difference in precision between UML and the other two risk model types is less clear as we can see from the overlapping boxplots on the right of the figure. There is a significant difference in recall between tabular and UML representations, but the difference between CORAS and UML is not significant.




(b) Complex questions

The difference in either precision or recall between tabular and CORAS notations are significant in favor of the former.

Figure 1: Participants' precision and recall by modeling notation

Table 8: RQ2: Summary of the findings

(a) Precision

| Finding | Statistical test results |
|---|---|
| Tabular $\gtrsim$ UML | $p_{TOST_{MW}} = 0.04$ |
| Tabular > CORAS | $p_{MW} = 0.0009$ ($p_{KW} = 0.002$) |

(b) Recall

| Finding | Statistical test results |
|---|---|
| Tabular > UML | $p_{MW} = 0.004$ ($p_{KW} = 0.008$) |
| Tabular > CORAS | $p_{MW} = 1.4 \cdot 10^{-5}$ ($p_{KW} = 6.6 \cdot 10^{-5}$) |
| UML > CORAS | $p_{MW} = 0.04$ |

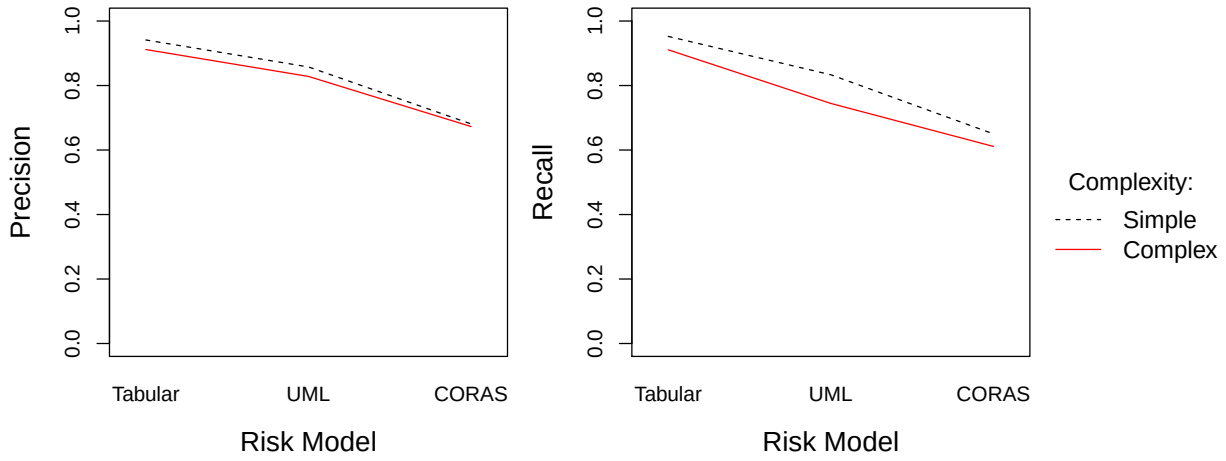



Figure 2: Interaction among risk modeling notation and task complexity

MW test assumes equality of variances we used Levene's test to check this assumption. For precision and recall of the responses from tabular and CORAS groups and for recall of the responses from tabular and UML groups the Levene's test revealed that the samples do not have equal variances and, therefore, we have to use KW test (we report the results of both MW and KW tests to provide comparison with other findings). Table 8 summarizes the findings of a post-hoc MW test with Bonferroni correction ($\alpha = 0.05/3 = 0.017$) and TOST with MW test and $\delta = 0.12$.

The availability of textual labels helps to give more precise responses as the results showed that tabular and UML models are equivalent w.r.t. precision of responses whilst tabular model showed significantly better precision than the graphical model. For recall tabular is significantly better than the other two notations, but the difference between UML and CORAS is unclear. Thus, we reject the null hypothesis $H2_0$ for precision, whilst the question remains open for the recall.

### *6.3. Interaction and co-factor analysis*

Figure 2 illustrates the lack of interaction between task complexity and modeling notation. It also illustrates that the difference between simple and complex questions is almost negligible.

We used the permutation test for two-way ANOVA to investigate the possible interaction between independent and dependent variables with several co-factors: participants' education degree, level of English, working experience, the level of participants' knowledge of architectural and system specification and modeling, security architecture and technology, risk assessment, graphical modeling languages, online banking. There was no statistically significant interaction between risk modeling type, dependent variables and any co-factor.

Table 9: Post-task questionnaire results
For all modeling notations participants agreed that settings were clear, tasks were reasonable, and documentation was appropriate. Participants who used CORAS and UML models experienced problems in understanding and answering comprehension questions due to the problem discussed in Sec. 4. Scale from 1 (strongly disagree) to 5 (strongly agree).

| | Tabular | | | UML | | | CORAS | | |
|---|---|---|---|---|---|---|---|---|---|
| Q# | mean | med | sd | mean | med | sd | mean | med | sd |
| Q1 | 4.48 | 5.00 | 0.93 | 4.20 | 4.50 | 0.95 | 4.15 | 4.50 | 1.09 |
| Q2 | 4.19 | 4.00 | 0.81 | 4.00 | 4.00 | 0.65 | 3.65 | 4.00 | 1.18 |
| Q3 | 4.57 | 5.00 | 0.60 | 4.05 | 4.00 | 1.05 | 3.85 | 4.00 | 1.04 |
| Q4 | 3.90 | 4.00 | 1.00 | 3.35 | 4.00 | 1.18 | 3.30 | 3.50 | 1.13 |
| Q5 | 4.14 | 4.00 | 0.85 | 3.45 | 4.00 | 1.10 | 3.15 | 3.00 | 1.04 |
| Q6 | 4.33 | 5.00 | 1.20 | 3.75 | 4.00 | 1.16 | 3.50 | 4.00 | 1.15 |
| Q7 | 4.33 | 5.00 | 1.20 | 3.75 | 4.00 | 1.16 | 3.50 | 4.00 | 1.15 |
| Q8 | 4.57 | 5.00 | 0.81 | 4.30 | 4.00 | 0.57 | 4.10 | 4.00 | 0.72 |
| Q9 | Yes (71%) / No (29%) | | | Yes (30%) / No (70%) | | | Yes (70%) / No (30%) | | |
| Q10 | Yes (95%) / No (5%) | | | Yes (55%) / No (45%) | | | Yes (75%) / No (25%) | | |

### *6.4. Post-task questionnaire*

We asked participants to provide their feedback on experiment execution with post-task questionnaire. Table 9 presents descriptive statistics of participants' feedback. Responses are on a five-item Likert scale from 1 (strongly disagree) to 5 (strongly agree).

For all three risk modeling notations participants concluded that the time allocated to complete the task was enough ($Q1$). They agreed that the objectives of the study ($Q2$) and the task ($Q3$) were clear. In general, participants found the comprehension questions to be clear ($Q4$) and they did not experience difficulty in answering the comprehension questions ($Q5$). As expected, the participants with a lower result comprehension were less confident in their responses to these questions (see discussion in Sec. 4). Overall, the participants did not experienced significant difficulties in understanding ($Q6$) and using electronic versions ($Q7$) of risk model tables or diagrams. The online survey tool was also easy to use ($Q8$).

## 7. Discussion and Conclusions

We can summarize the finding of our study as follows:

**RQ1:** *What is the effect of task complexity on participants' actual comprehension of information presented in risk models?*

The results of our study with professionals showed small difference (3-6%) in the F-measure of participants' responses to simple and complex questions with all three risk modeling notations, but not statistically significant. For the tabular notation we could actually establish statistical equivalence between the performance on complex and simple questions using a proportional equivalence test.

**RQ2:** *Does the availability of textual labels improve participants' effectiveness in extracting correct information about security risks?*

Tables better support participants in recalling correct information about security risks in comparison to the diagrams with icon-based notation. The UML-like notation seems to be an enhanced version of CORAS representation that helps participants to find information. Therefore, the availability of textual labels helps to elicit better responses.

A possible explanation of the better comprehension with tables could be the ability to perform computer-aided searches and copy&paste information from the documents to the response form. We asked our participants whether they used these possibilities. Several participants using tables used searching/filtering as well as copy&paste. However, most of participants who used CORAS model also used search and copy&paste. In contrast, a third of the participants using UML model searched information in PDF and around half of them copy&paste. Therefore, computer aided searches cannot explain the difference in comprehensibility. A possibility could be that participants with CORAS might search for the exact titles of elements (i.e. information cues) as they could not map element types because CORAS used icons. In contrast, the participants with UML had the textual labels with elements' types, were able to locate by themselves the elements mentioned in the question and, thus, did not need to use search. This could be the subject of an eye-tracking experiment.

Our findings are obviously limited by the set-up of our study and there might be other graphical notation for risk modeling that support better understanding of security risks. Previous studies (Grøndahl et al. 2011; Hogganvik and Stølen 2005; Massacci and Paci 2012) give us some confidence that the selected notations are the best ones available at present.

For example, there might be other feature of graphical notations that our experiment have not captured yet. The memorization of information about security risks might be such feature: users might not have models available at all times and it might be differences in outcomes when they have to answers questions by recalling it from memory.

In spite of such disclaimers, a clear picture emerges from the empirical experiments from our team and other researchers aiming to determine the empirical difference between tabular and graphical notations. It is summarized in Table A.10 in the appendix.

At present, and in spite of the large academic research into graphical based notations for (security, risk, and other kind of) requirements, *diagrams do not actually help* in term of either design or comprehension. It is thus likely that industry will continue to use tables and text, and rightly so.

## References


Abrahao, S., C. Gravino, E. Insfran, G. Scanniello, and G. 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), 327–342.

Acar, Y., M. Backes, S. Fahl, S. Garfinkel, D. Kim, M. L. Mazurek, and C. Stransky (2017). Comparing the usability of cryptographic apis. In Proceedings of the 38th IEEE Symposium on Security and Privacy.

Agarwal, R., P. De, and A. P. Sinha (1999). Comprehending object and process models: An empirical study. IEEE Trans. Soft. Eng. 25(4), 541–556.

de la Vara, J. L., B. Marín, G. Giachetti, and C. Ayora (2016). Do models improve the understanding of safety compliance needs?: Insights from a pilot experiment. In Proc. of ESEM 2016, pp. 32. ACM.

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

Food and Drug Administration (2001). Guidance for industry: Statistical approaches to establishing bioequivalence.

Giacalone, M., F. Paci, R. Mammoliti, R. Perugino, F. Massacci, and C. Selli (2014). Security triage: an industrial case study on the effectiveness of a lean methodology to identify security requirements. In Proc. of ESEM 2014, pp. 24. ACM.

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

Grøndahl, I. H., M. S. Lund, and K. Stølen (2011). Reducing the effort to comprehend risk models: Text labels are often preferred over graphical means. Risk Analysis 31(11), 1813–1831.

Hadar, I., I. Reinhartz-Berger, T. Kuflik, A. Perini, F. Ricca, and A. 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), 1823–1843.

Heijstek, W., T. Kühne, and M. R. Chaudron (2011). Experimental analysis of textual and graphical representations for software architecture design. In Proc. of ESEM 2011, pp. 167–176. IEEE.

Hernan, S., S. Lambert, T. Ostwald, and A. Shostack (2006). Threat modeling-uncover security design flaws using the stride approach. MSDN Magazine-Louisville, 68–75.

Hogganvik, I. and K. Stølen (2005). On the comprehension of security risk scenarios. In Proc. of IWPC 2005, pp. 115–124. IEEE.

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

Kaczmarek, M., A. Bock, and M. Heß (2015). On the explanatory capabilities of enterprise modeling approaches. In Proc. of EEWC 2015, pp. 128–143. Springer.

Labunets, K., F. Massacci, and F. Paci (2017). On the equivalence between graphical and tabular representations for security risk assessment. In Proc. of REFSQ 2017, pp. 191–208. Springer.

Labunets, K., F. Massacci, F. Paci, S. Marczak, and F. M. de Oliveira (2017). Model comprehension for security risk assessment: an empirical comparison of tabular vs. graphical representations. Empir. Soft. Eng., 1–40.

Labunets, K., F. Massacci, F. Paci, and L. M. S. Tran (2013). An Experimental Comparison of Two Risk-Based Security Methods. In Proc. of ESEM 2013, pp. 163–172. IEEE.

Labunets, K., F. Paci, F. Massacci, and R. Ruprai (2014). An experiment on comparing textual vs. visual industrial methods for security risk assessment. In Proc. of EMPIRE at RE 2014, pp. 28–35. IEEE.

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

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

Matulevičius, R. (2014). Model comprehension and stakeholder appropriateness of security risk-oriented modelling languages. In Proc. of BPMDS 2014, pp. 332–347. Springer.

Mayer, N., A. Rifaut, and E. Dubois (2005). Towards a risk-based security requirements engineering framework. In Proc. of REFSQ 2005, Volume 5.

Mellado, D., E. Fernández-Medina, and M. Piattini (2006). Applying a security requirements engineering process. In Proc. of ESORICS 2006, pp. 192–206. Springer.

Mouratidis, H. and P. Giorgini (2007). Secure tropos: a security-oriented extension of the tropos methodology. Int. J. Softw. Eng. Know. Eng. 17(02), 285–309.

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

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

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

Scanniello, G., C. Gravino, M. Genero, J. Cruz-Lemus, and G. Tortora (2014). On the impact of uml analysis models on source-code comprehensibility and modifiability. ACM Trans. Soft. Eng. Meth. 23(2), 13.

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

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

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

Sharafi, Z., A. Marchetto, A. Susi, G. Antoniol, and Y.-G. Guéhéneuc (2013). An empirical study on the efficiency of graphical vs. textual representations in requirements comprehension. In Proc. of ICPC 2013, pp. 33–42. IEEE.

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

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

Stålhane, T. and G. 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), 1–24.

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

Van Lamsweerde, A. (2001). Goal-oriented requirements engineering: A guided tour. In Proc. of RE 2001, pp. 249–262. IEEE.
Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. Decision. Sci. 22(2), 219–240.
Wood, R. E. (1986). Task complexity: Definition of the construct. Organ. Behav. Hum. Dec. 37(1), 60–82.

## Appendix A. Appendix

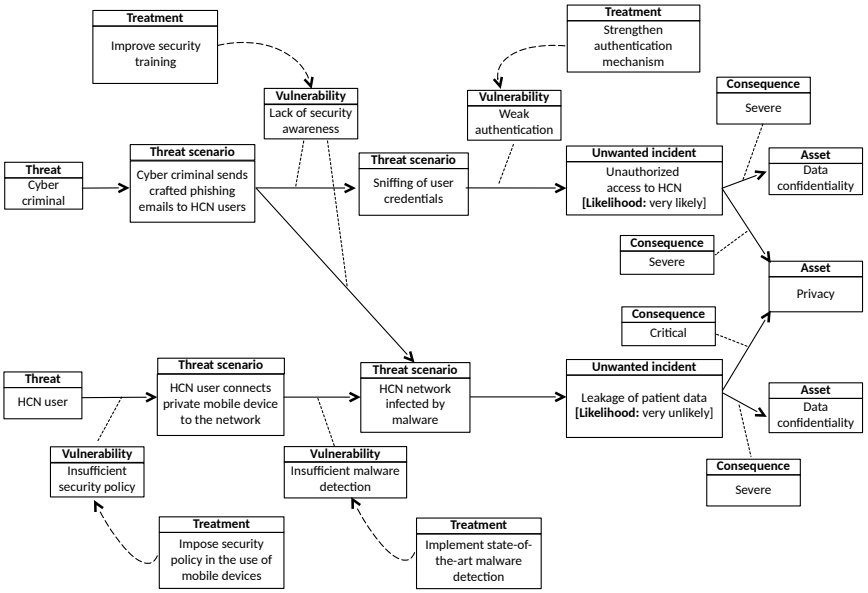
A full replication guide is available at https://securitylab.disi.unitn.it/doku.php?id=online-comprehensibility-exp-2016.

Table A.10: A summary of experiments on tabular vs graphical notations
Several experiments compared the effectiveness of textual descriptions in natural language (NL), tables (T), diagrams (D), and diagrams with textual labels (D+L). Participants were students (BSc, MSc, PhD) or professionals (Profs). Participants designed models assessed later by domain experts (Design) or answered questions about models to test their comprehensibility (Compr.). Researchers measured perceived ease of use or actual differences in outcomes. Diagrams do not seem to help; so industry is presently right in using tables and text.

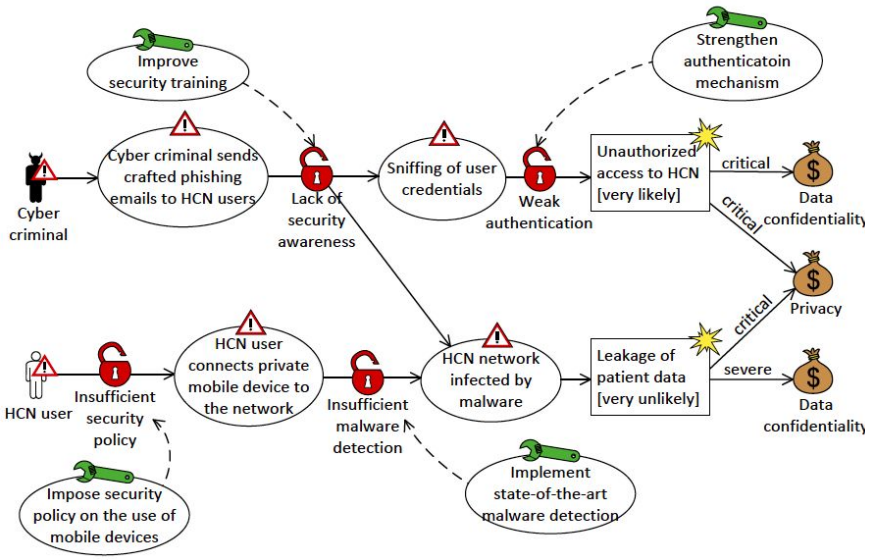
| Topic | Type | Comparison | #Subjects | Findings | Refs |
|---|---|---|---|---|---|
| Design | Exper. | CORAS (D), Problem Frames (D), Secure Tropos (D), si* (D) | 13 MSc, 36 Profs | CORAS perceived better than others methods for the analysis | Massacci and Paci (2012) |
| Design | Exper. | SREP (T), CORAS (D) | 28 MSc | Diagrams perceived better than Tables, Tables actually better on controls identification, Diagrams actually better on threats | Labunets et al. (2013) |
| Design | Exper. | Eurocontrol (T), CORAS (D) | 29 MSc | Diagrams perceived better than Tables, no actual difference | Labunets et al. (2014, 2017) |
| Design | Exper | Eurocontrol (T), CORAS (D) | 83 MSc | Tables & Diagrams actually equivalent, No difference in perception | Labunets et al. (2017) |
| Compr. | Survey | CORAS (D), CORAS+ (D+L) | 33 Profs | Textual labels help | Grøndahl et al. (2011) |
| Compr. | Exper | Description (NL), UML (D+L) | 21 MSc, 14 Studs, 12 Profs | Diagrams perceived better, but Text actually better, no difference on topological questions | Heijstek et al. (2011) |
| Compr. | Exper | Description (NL), TROPOS (D) | 2 BSc, 11 MSc, 15 PhD | No difference between Text & Diagrams, Diagrams took longer | Sharafi et al. (2013) |
| Compr. | Exper | DOC/EN (NL), UML (D+L) | 15 BSc | No difference between Text and Diagrams | de la Vara et al. (2016) |
| Compr. | Exper | NIST (T), CORAS (D) | 104 MSc, 27 BSc | Tables actually better than Diagrams, Complex questions have poorer recall in one study but no difference in a second study | Labunets et al. (2017) |
| Compr. | Exper | NIST (T), CORAS (D), UML (D+L) | 63 Profs | Tables actually better than Diagrams but textual labels help, No difference due to questions' complexity | here |

| 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<br>2. Weak authentication | Unauthorized access to HCN | Data confidentiality | Very likely | Severe | 1. Improve security training.<br>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<br>2. Weak authentication | Unauthorized access to HCN | Privacy | Very likely | Severe | 1. Improve security training.<br>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<br>2. Insufficient malware detection | Leakage of patient data | Privacy | Very unlikely | Critical | 1. Impose security policy on the use of mobile devices.<br>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<br>2. Insufficient malware detection | Leakage of patient data | Data confidentiality | Very unlikely | Severe | 1. Impose security policy on the use of mobile devices.<br>2. Implement state-of-the-art malware detection. |

(a) NIST table row entries




(b) UML diagram




(c) CORAS diagram

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

Table A.11: Precision and recall by questions

The biggest difference ($\geq 0.2$) in precision was observed for Q1, Q7 and Q10 and in recall for Q1, Q3, Q5, Q7, Q10 and Q12 between tabular and CORAS models in favor of the former. Between tabular and UML models the most significant difference was observed in precision and recall of Q12 in favor of tabular notation. The missing responses in column "*#obs.*" can be caused by task termination forced by SurveyGizmo due to time limit.

| Q# | Comp- | Tabular | | | | UML | | | | CORAS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | lexity | #obs. | mean | med. | sd | #obs. | mean | med. | sd | #obs. | mean | med. | sd |
| Precision | | | | | | | | | | | | | |
| Q1 | 2 | 21 | 0.95 | 1.00 | 0.22 | 20 | 0.80 | 1.00 | 0.41 | 20 | 0.45 | 0.00 | 0.51 |
| Q2 | 4 | 21 | 0.95 | 1.00 | 0.15 | 20 | 0.90 | 1.00 | 0.26 | 19 | 0.92 | 1.00 | 0.25 |
| Q3 | 2 | 21 | 0.95 | 1.00 | 0.16 | 20 | 0.97 | 1.00 | 0.15 | 20 | 0.80 | 1.00 | 0.41 |
| Q5 | 6 | 21 | 0.93 | 1.00 | 0.23 | 18 | 0.88 | 1.00 | 0.29 | 19 | 0.83 | 1.00 | 0.37 |
| Q7 | 4 | 21 | 0.88 | 1.00 | 0.31 | 20 | 0.75 | 1.00 | 0.44 | 20 | 0.60 | 1.00 | 0.50 |
| Q10 | 4 | 21 | 0.81 | 1.00 | 0.40 | 19 | 0.68 | 1.00 | 0.45 | 19 | 0.42 | 0.00 | 0.51 |
| Q12 | 6 | 21 | 0.90 | 1.00 | 0.27 | 19 | 0.68 | 1.00 | 0.48 | 19 | 0.47 | 0.00 | 0.51 |
| Overall | | 147 | 0.94 | 1.00 | 0.22 | 136 | 0.80 | 1.00 | 0.39 | 136 | 0.69 | 1.00 | 0.46 |
| Recall | | | | | | | | | | | | | |
| Q1 | 2 | 21 | 0.95 | 1.00 | 0.22 | 20 | 0.80 | 1.00 | 0.41 | 20 | 0.45 | 0.00 | 0.51 |
| Q2 | 4 | 21 | 0.95 | 1.00 | 0.15 | 20 | 0.85 | 1.00 | 0.29 | 19 | 0.79 | 1.00 | 0.30 |
| Q3 | 2 | 21 | 0.98 | 1.00 | 0.11 | 20 | 0.92 | 1.00 | 0.18 | 20 | 0.75 | 1.00 | 0.41 |
| Q5 | 6 | 21 | 0.90 | 1.00 | 0.26 | 18 | 0.76 | 1.00 | 0.34 | 19 | 0.67 | 0.75 | 0.40 |
| Q7 | 4 | 21 | 0.90 | 1.00 | 0.30 | 20 | 0.75 | 1.00 | 0.44 | 20 | 0.60 | 1.00 | 0.50 |
| Q10 | 4 | 21 | 0.81 | 1.00 | 0.40 | 19 | 0.74 | 1.00 | 0.45 | 19 | 0.42 | 0.00 | 0.51 |
| Q12 | 6 | 21 | 0.90 | 1.00 | 0.30 | 19 | 0.68 | 1.00 | 0.48 | 19 | 0.47 | 0.00 | 0.51 |
| Overall | | 147 | 0.89 | 1.00 | 0.28 | 136 | 0.73 | 1.00 | 0.38 | 136 | 0.60 | 0.75 | 0.44 |