Second evaluation report and method selection guidelines

Document information	
Project Title	EMFASE
Project Number	E.02.32
Project Manager	University of Trento
Deliverable Name	Second evaluation report and method selection guidelines
Deliverable ID	D2.3
Edition	00.01.00
Template Version	03.00.00
Task contributors	
Deep Blue; University of Trento; SINTEF	

Please complete the advanced properties of the document

Abstract

The main objective of EMFASE WP2 is to evaluate principles and methods concerned with risk assessment. These methods and principles are empirically evaluated to produce selection guidelines. This deliverable presents the second version of the EMFASE empirical evaluation framework and how it has been applied to different experiments. It summarizes the results obtained from the empirical studies conducted so far in, the lessons learnt and the guidelines for selecting Security Risk Assessment methods.

Authoring & Approval

Prepared By - *Authors of the document.*		
Name & Company	Position & Title	Date
Martina Ragosta / Deep Blue srl.	Project Contributor	09/02/2016
Alessandra Tedeschi / Deep Blue srl.	WP2 leader	15/02/2016
Katsiaryna Labunets / UNITN	Project Contributor	09/03/2016

Reviewed By - *Reviewers internal to the project.*		
Name & Company	Position & Title	Date
Rainer Koelle / SJU	WP-E Officer	14/03/2016

Approved for submission to the SJU By - *Representatives of the company involved in the project.*		
Name & Company	Position & Title	Date
Fabio Massacci /UNITN	Project Manager	14/03/2016

Rejected By - *Representatives of the company involved in the project.*		
Name & Company	Position & Title	Date
<Name / Company>	<Position / Title>	<DD/MM/YYYY>

Rational for rejection
None.

Document History

Edition	Date	Status	Author	Justification
00.00.01	09/02/2016	First draft	Martina Ragosta	New Document
00.00.02	15/02/2016	Working draft	Alessandra Tedeschi, Marta Ceccaroni	Internal review; ch. 1-5
00.00.03	04/03/2016	Draft revision	Rainer Koelle	Review ch.1-5
00.00.04	09/03/2016	Internal release	Martina Ragosta, Alessandra Tedeschi, Marta Ceccaroni, Katsiaryna Labunets	Project Officer's comments addressed. Preliminary version of the whole document
00.00.05	11/03/2016	Preliminary release	Elisa Chiarani and Fabio Massacci	Quality check and PM approvement for submission to the SJU
00.01.00	14/03/2016	Final release	Rainer Koelle	Final review

Intellectual Property Rights (foreground)

This deliverable consists of foreground owned by one or several Members or their Affiliates.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Purpose of the document

The main objective of EMFASE WP2 is to provide support to decision makers for selection of risk assessment methods for security in the ATM domain. This support takes the form of guidelines for how to select the risk assessment method best suited for the particular situation it is to be used and the role of the stakeholders to use it. These guidelines will be developed for evaluating risk assessment methods adopted in practice based on criteria that originate from end-user goals and relevant ATM standards. To define these guidelines, it is needed to evaluate risk assessment methods that have been carefully chosen as the objects of study, and the application scenarios and the assessment studies study designs based on them. The empirical evaluation is accomplished through case studies and/or controlled experiments as prescribed by the empirical evaluation framework developed in [3] and refined in [5]. This document presents the second version of the EMFASE empirical evaluation framework and how it has been applied to different experiments. It summarizes the results obtained from the empirical studies conducted so far in, lessons learnt and SRA method selection guidlines. More specifically the document is structured as follows:

- **Section** 2 presents the refined empirical evaluation framework and contain a detailed methodological and procedural description of each controlled experiment conducted so far in about risk models comprehensibility.
- **Sections** 4 **to** 8 provide a detailed experiments execution of the five experiments regarding risk models comprehensibility. They include also the achieved results, some additional analyses to provide causal explanation of the experiments findings will be presented in D3.2.
- **Section** 9 describes lessons learnt from the comprehensibility experiments.
- **Section** 10 offers a final overview on the EMFASE evaluation results.

1.2 Intended readership

As stated in Section 1.1, D2.3 is mainly an internal working document for EMFASE. Thus, intended readers of this document are primarily the EMFASE project partners and the EUROCONTROL Project Officers that have to agree on the framework and on the revised guidelines. Accordingly, this document is meant to be used by the members of the project EMFASE as it provides information about the controlled experiments that will serve as a read hearing throughout the project. In particular, the content of the document will be used as input/feedback to the activities of WP1 in which the lessons learned from the actual evaluation designs and evaluations will be generalized and incorporated into the evaluation framework. Additionally, the phenomena observed in the evaluations will be used as input and further explained in the WP3 which will provide causal explanations of them. Other potential readers are generally all stakeholders within the ATM domain that need to take security into account in an operational area. More specifically, the document is of interest to all SESAR JU projects within the transversal areas of WP16 that are related to security management and risk assessment, in particular SESAR 16.06.02. For these stakeholders the document gives insight into some of ATM security risk assessment methods that could be relevant to apply or investigate further.

1.3 Inputs from other projects

The document does not make use of input from other projects, but the content is related to both SESAR 16.02.03 and SESAR 16.06.02 for what regards the SESAR SecRAM Security Risk Assessment Methodology. References to these projects are given in the relevant sections.

1.4 Acronyms and Terminology

Table 1: Acronyms and Terminology

Term	Definition
AE	Actual Effectiveness
ATM	Air Traffic Management
EMFASE	Empirical Framework for Security Design and Economic Trade-Off
HCN	Health Communication Network
RQ	Research Questions
SecRAM	Security Risk Assessment Methodology
SESAR	Single European Sky ATM Research Programme
SESAR Programme	The programme which defines the Research and Development activities and Projects for the SJU
SJU	SESAR Joint Undertaking (Agency of the European Commission)
SJU Work Programme	The programme which addresses all activities of the SESAR Joint Undertaking Agency
SRA	Security Risk Assessment

2 Refined Empirical Evaluation Framework

2.1 Best Practices on Empirical Methods

Security risk assessment (SRA) involves human interaction and communication, the use of methods and techniques, decision making based on risk documentation, and several other real life issues. Analytical research is often not sufficient for investigating such, sometimes complex, issues. Instead it may be necessary to conduct empirical research in order to gather empirical evidence and develop theories for the objects of study [15]. EMFASE is concerned with practitioners' use of SRA methods within the ATM domain, as well as the use of the risk assessment results by decision makers and other stakeholders. Which SRA techniques and activities are best suited for which needs, and why is that so? In conducting empirical studies and in developing the empirical framework, EMFASE makes use of established empirical research methods and best practices for how to conduct empirical studies. For an overview of relevant research methods and references to literature we refer to D1.2. EMFASE follows established guidelines and best practices for how to conduct and report empirical studies. Based on such guidelines we follow the research process of [15] as outlined in the following table. The process is mostly the same for all kinds of empirical studies, although it is often conducted more iteratively for more flexible research like case studies

Table 2: Case study research process

Step	Activity
1	Empirical study design: Objectives are defined and the empirical study is planned
2	Preparation for data collection: Procedures and protocols for data collection are defined
3	Collecting evidence: Execution with data collection on the study subject
4	Analysis of collected data
5	Reporting

Step 1 involves defining the objectives of the study, i.e. what to achieve and which research questions to investigate. The subject of the study must also be specified. For EMFASE, the subject is typically the whole or parts of an SRA, including the people and interactions involved. Step 2 involves specifying the method for data collection, as well as the protocol for conducting the specific study. Step 3 is the collection of data during the execution of the study. Methods for data collection include interviews, observations, experiment output and archival data. The data analysis of Step 4 can be quantitative or qualitative. Quantitative analysis may involve analysis of statistics and correlations, as well as hypothesis testing and the development of predictive models. Qualitative analysis involves deriving conclusions from the gathered data, keeping a clear chain of evidence from the data to the conclusions that can be followed by the reader [15] [20]. The reporting of Step 5 shall document the findings of the study and serve as the main source for judging the quality of the study. A similar process for empirical research is presented in [10] where guidelines are proposed for each of the following steps: Experimental context, experimental design, conducting the experiment and data collection, analysis, presentation of results, and interpretation of results. These guidelines focus more on experimental studies than case study research, and therefore complement the case study guidelines in [15] for the process outlined in Table 2

2.2 Success Criteria for ATM Security Risk Assessment Methods

In order to enable an empirical evaluation and comparison of methods for security risk assessment we identified criteria with respect to which the methods shall be evaluated. There are of course many different parameters and aspects that can be considered for the classification and evaluation of methods for security risk assessment. In the EMFASE project, we derived the success criteria in close collaboration with ATM security stakeholders. In this section we present the identified criteria and how they relate to the MEM. For further details we refer to D1.2 [3]

2.2.1 Identified Success Criteria

The following table summarizes the main criteria reported by the professionals. We considered as the main identified criteria only the ones for which at least ten statements were made by the participants. Each of the criteria is explained in the next section, but we can observe here that while the main bulk of the statements fall into six main categories, the total share of other statements is significant. This indicates some spread in the opinions of the ATM stakeholders. Some of the less frequent statements (i.e. "Tool support", "Comparability of results" and "Comprehensibility of results") were considered as relevant by EMFASE security experts and thus introduced as well in the overall list of EMFASE success criteria that may be subject to empirical investigation.

Table 3: Occurrences of reported success criteria

Criterion	**N of Statements**
Clear steps in the process	28
Specific controls	24
Easy to use	19
Coverage of results	14
Tool support	13
Comparability of results	10
Others	51
Total	**159**

In our classification and evaluation of security risk assessment methods we take into account all additional support that comes with each method. Some security risk assessment methods come with repositories of assets and controls, while other methods come with tools for risk modelling. Guided by the identified criteria, EMFASE security experts identified further method features or artefacts that could contribute to fulfill the criteria. Some of these correspond to criteria identified also by the ATM stakeholders. They are additional properties/features of security risk assessment methods that can contribute to support one or more of the six main criteria identified by the professionals. A more detailed description is presented in D1.1 [4].

In order to further structure the success criteria, EMFASE security experts aggregated the criteria and preliminarily categorized them into four main categories:

- **Process**: The steps for conducting the SRA.
- **Presentation**: The means for specifying and documenting the SRA results.
- **Results**: The output from the SRA.
- **Supporting material**: Any support that comes with an SRA method, such as tools and catalogues.

As a result the following classification is the scheme for supporting the method evaluation in EMFASE.

Table 4: EMFASE success criteria categories

Process	**Presentation**	**Results**	**Supporting material**

Clear steps in the process;
Time effective;
Holistic process;
Compliance with ISO/IEC standards
	Easy to use;
	Help to identify threats;
	Visualization;
	Systematic listing;
	Comprehensibility of method outcomes;
	Applicable to different domains;
	Evolution support;
	Well-defined terminology
		Specific controls;
		Coverage of results;
		Comparability of results
			Tool support;
			Catalogue of threats and security controls;
			Worked examples;
			Documentation templates;
			Modelling support;
			Practical guidelines;
			Assessment techniques

3 Experiments on comprehensibility of risk models

Among method's success criteria (see Table 4) we identified category "Comprehensibility of method outcomes". We reviewed the experts' statements that were included in this category and discuss them [11] in order to understand the role of comprehensibility in a security risk assessment.

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

As we have seen, the effective communication of the results with different stakeholders is an important factor for the success of security risk assessment. However, which risk model representation is more comprehensive for the stakeholders?

To design our comprehensibility task we revised existing works investigating models comprehensibility in requirements [7, 16] and data modeling [2, 13]. We found out that comprehensibility questions in these studies aim to test the ability of the user to identify 1) a risk element of a specific type that is in relationship with another element of a different type and 2) a risk element of a specific type that has multiple relationships with other elements of a different type. We used both approaches to formulate questions for our comprehensibility task.

3.1 Overview

In this section we describe the empirical studies which have been conducted in EMFASE following the empirical protocol described in the previous section. The following picture shows the timeline of these studies.

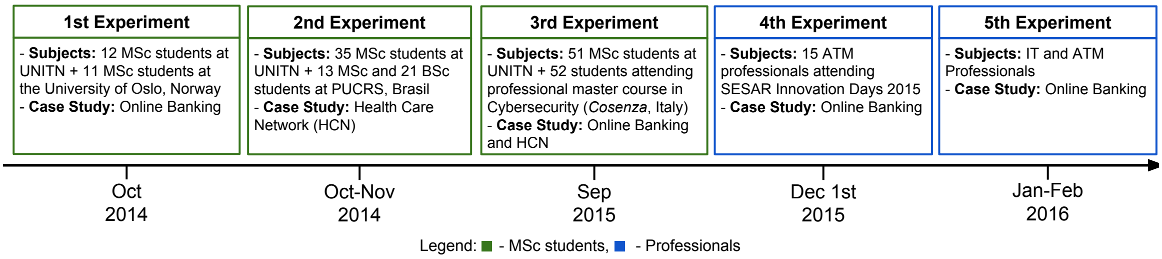

1st Experiment
- Subjects: 12 MSc students at UNITN + 11 MSc students at the University of Oslo, Norway
- Case Study: Online Banking
2nd Experiment
- Subjects: 35 MSc students at UNITN + 13 MSc and 21 BSc students at PUCRS, Brasil
- Case Study: Health Care Network (HCN)
3rd Experiment
- Subjects: 51 MSc students at UNITN + 52 students attending professional master course in Cybersecurity (Cosenza, Italy)
- Case Study: Online Banking and HCN
4th Experiment
- Subjects: 15 ATM professionals attending SESAR Innovation Days 2015
- Case Study: Online Banking
5th Experiment
- Subjects: IT and ATM Professionals
- Case Study: Online Banking
Oct 2014
Oct-Nov 2014
Sep 2015
Dec 1st 2015
Jan-Feb 2016
Legend: 🟩 - MSc students, 🟦 - Professionals

Figure 1: Empirical studies timeline

While the first three Experiments have been conducted with MSc students then the two other ones with ATM and IT Professionals (Experiment 3 and 4). Albeit with some variations, the experiments focus on the comprehensibility of Security Risk Assessment methodologies and the research method is described in the following subsection.

3.2 Research method

We conducted the experiments by following established methods for empirical research [18, 20]. Our objective was explanatory, i.e. to seek an explanation in the form of a causal relationship [15] in order to investigate the effect of the risk model format on the comprehensibility of risk models. We found out that comprehensibility questions in these studies aim to test the ability of the user to identify 1) a risk element of a specific type that is in relationship with another element of a different type and 2) a risk element of a specific type that has multiple relationships with other elements of a different type. We used both approaches to formulate questions for our comprehensibility task.

3.2.1 Task Complexity and Other Factors

Computing task complexity. We also take into consideration the complexity of the questions as it may be a significant factor for the risk model comprehensibility. To define it we rely upon the work of Wood [19], according to which task (or question) complexity is defined by the information cues that need to be processed and the number and complexity of the actions that need to be performed to accomplish the task:

- "Information cues are pieces of information about the attributes of stimulus objects" [19, p. 65]
- "The required acts for the creation of a defined product [output] can be described at any one of several levels of abstraction. . . " [19, p. 66]
- "Coordinative complexity refers to the nature of relationships between task input and task product. As the number of precedence relationships between acts increases, the knowledge and skill required will also increase. . . " [19, p. 68-69]

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

- *Information cues (IC)* describe some characteristics that help to identify the desired element of the model. They are identified by a noun. In the sentence "Which are the assets that can be harmed by the unwanted incident *Unauthorized access to Health Communication Network (HCN)*?" the part in italics is an information cue.
- *Required acts (A)* are judgment acts that require to select a subset of elements meeting some explicit or implicit criteria. For example, in "What is the *highest* consequence?" or "What are

the unwanted incidents that *can* occur?" the parts in italics are judgment criteria.

- *Relationships (R)* are relationship between a desired element and other elements of the model that must to be identified in order to find the desired element. They are identified by a verb. In the sentence "the assets that can *be harmed by*", the part in italics is a relationship.

To calculate the *complexity of question* i (QC_i) we use the following formula:

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

where IC_i is the set of information cues presented in the question i, R_i is the set of relationships that the subject needs to identify and A_i is the set of judgments to be perform over a set of elements, and by $|\ |$it is meant the cardinality of a set.

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

Other factors. Another possible confounding factor is the complexity of the particular execution of the experiment itself. Therefore, after the comprehension task we asked subjects to fill in a post-task questionnaire about their perception of the clarity of the questions and the overall settings and whether the risk model was easy to understand. The aim of the post-task questionnaire is to control the possible effects of the experimental settings on the results as done in previous studies [7, 1]. Table 6 reports the post-task questionnaire that we proposed to our subjects.

Variables. The *independent variable* of our study is the risk model representation which can take one of the values: tabular or graphical. The *dependent variable* is the level of comprehensibility which is measured by assessing the answers of the subjects to a series of comprehension questions about the content presented in the risk models. In the sequel, we will use the word 'task' when referring to the entire exercise of answering all questions. The answers to the questions were evaluated using information retrieval metrics that are widely adopted in the empirical software engineering community for the measurement of models comprehension [1, 7, 16]: *precision*, *recall*, and their harmonic combination, the *F-measure*. Precision represents the correctness of given responses for the question, and recall represents the completeness of the responses. They are calculated as follows:

$$\begin{aligned}
precision_{m,s,i} &= \frac{|answer_{m,s,i} \cap correct_{i,m}|}{|answer_{m,s,i}|}, && (2)\\
recall_{m,s,i} &= \frac{|answer_{m,s,i} \cap correct_{i,m}|}{|correct_{i,m}|}, && (3)\\
F_{m,s,i} &= 2 * \frac{precision_{m,s,i} * recall_{m,s,i}}{precision_{m,s,i} + recall_{m,s,i}}, && (4)\\
& && (5)
\end{aligned}$$

where $answer_{m,s,i}$ is the set of answers given by subject s to question i when looking at model m; $correct_{i,m}$ is the set of correct responses to question i (which may vary with the Scenario and the Subexperiment considered).

As we have several questions for each subject, we used the mean of Precision, Recall and F-measures of all questions average measures for each subject. Further in the paper, if not specified differently, by average precision (or average recall, of average F-measure) we use the mean of the precisions (recalls, F-measures respectively) on all questions for each subject:

$$\bar{P}_{m,s} = \frac{\sum_{i=1}^{N_{questions}} precision_{m,s,i}}{N_{questions}}, \tag{6}$$

$$\bar{R}_{m,s} = \frac{\sum_{i=1}^{N_{questions}} recall_{m,s,i}}{N_{questions}}. \tag{7}$$

$$\bar{F}_{m,s} = \frac{\sum_{i=1}^{N_{questions}} F_{m,s,i}}{N_{questions}}, \tag{8}$$

$$\tag{9}$$

Precision, recall and F-measure values (and obviously their averages) range from 0 to 1. If a subject shows an average F-measure $\bar{F}_{m,s,e}$ value close to 1 means that the subject showed a good comprehension of the model m (during the session e).

To further check the robustness of the statistics, we use the aggregated F-measure calculated as following:

$$tP_{m,s} = \frac{\sum_{i=1}^{N_{questions}} |answer_{m,s,i} \cap correct_{i,m}|}{\sum_{i=1}^{N_{questions}} |answer_{m,s,i}|}, \tag{10}$$

$$tR_{m,s} = \frac{\sum_{i=1}^{N_{questions}} |answer_{m,s,i} \cap correct_{i,m}|}{\sum_{i=1}^{N_{questions}} |correct_{i,m}|}, \tag{11}$$

$$tF_{m,s} = 2 * \frac{tP_{m,s} * tR_{m,s}}{tP_{m,s} + tR_{m,s}}. \tag{12}$$

3.3 Analysis Procedure

We test the null hypothesis H_0 (about no difference in the actual comprehensibility between two types of risk models) we select different statistical tests depending on the design type and the assumption on normal distribution of the samples. Table 5 (shortened version of the Table 37.1 from [8, Chap. 37]) gives an overview of how we select statistical test in our empirical studies. For example, in Experiment 1 we could use either unpaired t-test or its non-parametric analog, Mann-Whitney test. Because the subjects of Experiment 1 applied only one of two treatments. However, in Experiment 3 the subjects used both treatments to two different application scenario and, therefore, we could use either paired t-test or Wilcoxon test. The actual statistical test being used will be reported in the sections describing each individual study. We tested the normality of the data distribution using the Shapiro-Wilk test. For all statistical tests we set the significance level $\alpha = 0.05$. To calculate the statistical power of the test for a null hypothesis acceptance we adopt 20% as a threshold for β (Type-II error) and use `G*Power 3` for the actual calculation [6].

We investigate the effect of the task complexity on actual comprehensibility using interaction plots. To further validate the interaction of the independent variable and task complexity we use

Table 5: Statistical Tests Selection

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

contingency tables for the number of subjects that achieved a F-measure above the median F-measure of the responses to low complexity and high complexity questions provided using tabular or graphical risk models.

To find whether the interaction between these variables is statistically significant we perform Fisher's test using the information from contingency tables [12, Sec. 17.4.1].

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

3.4 Experimental procedure

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

The different experiments may have specific settings varying from the above procedure. In this case these specific settings will be reported in the experiment description in the corresponding section.

4 Experiment 1: comprehensibility of risk models with MSC students

The main goal of this study was to investigate the comprehensibility of risk models expressed in two modeling approaches: graphical vs. tabular. We executed the study in the form of two controlled Subexperiments with MSc students. The first Subexperiment was conducted by UNITN with MSc students enrolled in the Security Engineering course at University of Trento, while the second one was conducted by SINTEF with MSc students of the Model Engineering course held at the University of Oslo. The comparison of two types of risk models was done using questionnaire about comprehensibility of specific aspects of risk models that were distributed to the participants.

4.1 Experiment execution

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

Procedure The subjects were given five minutes to answer a demographics and background questionnaire, after which they had 20 minutes to complete the comprehensibility questionnaire. Half of the subjects were given NIST risk tables and half of them were given CORAS diagrams. The tables and CORAS diagrams represented the same security risk documentation (i.e. same semantics) as derived from the Poste Italiane application scenario. After the completion of the comprehensibility questionnaire, the subjects had two minutes to fill in the post-task questionnaire reported in Table 6. The purpose of the post-task questionnaire was to control the possible effect of the experimental setting on the results. Table 6 contains the results of the post-task questionnaire that we distributed to the subjects. Questions Q1-Q8 included closed answers on a 5-point Likert scale: 0 – strongly agree, 1 – agree, 2 – not certain, 3 – disagree, and 4 – strongly disagree. Only question Q9 had "yes" and "no" answers.

Table 6: Post-task questionnaire

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

The ethical considerations were handled by informing the subjects in advance about the purpose of the study and how the gathered data is used by whom. Anonymity and confidentiality of personal

Table 7: Precision and recall by questions and risk model type – Experiment 1

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

data is guaranteed, and the processing and storage of the data is used only for the purposes of the study.

4.2 Results

4.2.1 Descriptive Statistics

Table 7 reports the descriptive statistics for precision and recall, both for the individual comprehension questions and overall. Overall, the two risk models had similar precision, but the tabular risk model demonstrated slightly higher recall than the graphical one. At the level of each comprehension question, the subjects who used tabular risk model showed higher precision and recall, with some exceptions. For question Q8 the two risk models demonstrated equal precision and recall. For the precision of Q3 and Q7 the graphical risk model outperformed the tabular one, while for Q2 the tabular model has slightly lower precision than the graphical model. For questions Q3 and Q7 the graphical risk model had higher recall that the tabular one.

4.2.2 Hypothesis Testing

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

Table 8 reports the mean, median, and standard deviation of precision, recall and F-measure by risk model type for each Subexperiment. To investigate the difference between two types of risk models we used Mann-Whitney test as our data is not normally distributed. The last two columns report Z statistics and p-value returned by the Mann-Whitney test.

The results of the test show that the subjects from Oslo demonstrated similar precision and recall using either graphical or tabular risk models. In contrast, the subjects from Trento showed better results using tabular risk model than using the graphical one. Overall, the subjects achieved equal

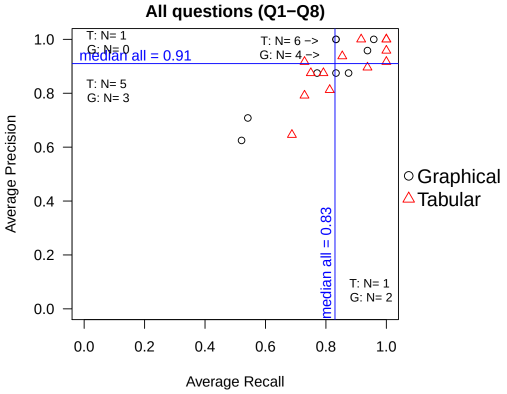


Figure 2: Distribution of Average Precision vs. Average Recall per subject by risk model type – Experiment 1

Table 8: Average precision, recall and F-measure by Subexperiments and risk model type – Experiment 1

	Tabular			Graphical			Mann-Whitney	
	Mean	Med.	sd	Mean	Med.	sd	Z	p-value
Precision								
SINTEF	0.90	0.93	0.09	0.94	0.96	0.06	0.83	0.40
UNITN	0.91	0.90	0.07	0.80	0.79	0.17	-1.16	0.25
Overall	0.90	0.92	0.08	0.88	0.88	0.13	-0.17	0.86
Recall								
SINTEF	0.85	0.83	0.09	0.87	0.88	0.08	0.55	0.58
UNITN	0.89	0.94	0.12	0.68	0.69	0.17	-1.53	0.13
Overall	0.87	0.85	0.11	0.79	0.83	0.16	-0.87	0.38
F-measure								
SINTEF	0.87	0.85	0.08	0.91	0.91	0.06	0.82	0.41
UNITN	0.90	0.92	0.10	0.74	0.73	0.17	-1.52	0.13
Overall	0.89	0.89	0.09	0.83	0.88	0.14	-0.60	0.55

level of actual comprehension of security risks (which we measured as F-measure) using both tabular and graphical risk models. The Mann-Whitney test did not reveal any significant difference between the two risk models for all Subexperiments and dependent variables. Thus, we cannot reject the null hypothesis H_0 for the scenario and circumstances studied in this experiment.

4.2.3 Post-task Questionnaire

We used the responses to the post-task questionnaire to control the possible effect of the experiment settings on the results. Table 9 reports mean, median and standard deviation of the responses by risk model type. The responses are on a 5-item Likert scale from 0 (strongly agree) to 4 (strongly disagree). Overall, all subjects—regardless the type of risk model used—conclude that the settings were clear, the task was reasonable and the materials were clear and sufficient.

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

Table 9: Post-task questionnaire results – Experiment 1

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

4.2.4 Co-factor analysis

To test the possible effect of the co-factors on the dependent variables we used the two-way ANOVA, which is robust in case of violation of normality assumption, and is widely accepted in the literature for co-factor analysis [7, 17, 14]. We considered co-factors like work experience, level of expertise in security, modeling languages, as well as in the domain of online banking. Only one subject reported his knowledge in modeling languages as "novice". Therefore, we merged this category with the category "beginner". Another subject reported his knowledge of the online banking domain as "proficient user", and therefore, we merged this category with the "competent user" category.

The results of two-way ANOVA revealed only one statistically significant interaction. There is an interaction between the Subexperiments (Trento and Oslo) and risk model type with the effect on the F-measure, and this is statistically significant according to the results of two-way ANOVA which returned p-value = 0.047). The F-measure results presented in Table 8, comparing the findings from Trento and Oslo, clearly illustrate this effect. The two-way ANOVA also revealed a statistically significant effect of the subjects' level of expertise in modeling languages on the recall (p-value = 0.03). The results show that the subjects with higher level of expertise in modeling languages (e.g., "proficient user") provided more complete responses (median recall is 0.94) than the subjects with average or low level of expertise (median recall is 0.82).

5 Experiment 2: comprehensibility of risk models with MSc and BSc students

As said, the aim of the experiments is to test and compare the comprehensibility of two different methods to display risk models, varying, although slightly, the subjects of the tests, the complexity of the questions, and the design of the tests in order to spot possible unwanted correlations between the sample used and the results found and validate and generalise the result.

With this aim in mind, Experiment 2 is designed as a *between-subject* experiment, where each examined subject has either a tabular or a graphical model to analyse. Moreover, the test is single factor (risk modeling notation) but double treatment (graphical/tabular). Results are analysed also taking into account the Complexity level of the questions evaluated by the number of logical connections, clues and judgments contained in the questions.

Some basic analysis of the demographical features and the answers to the post-task questionnaire of the sample examined are first taken into account, in order to compare the samples with the other experiments and stress possible influences on the result.

5.1 Experiment execution

Experiment was divided in three different subexperiments enrolling both BSc and MSc students from different universities. The first subexperiment was conducted by UNITN with 35 MSc students in Computer Science at University of Trento, in the fall semester of 2014 as part of the University course. The second and third subexperiments were conducted at the PUCRS University in Porto Alegre (Brasil) with 13 MSc and 27 BSc students of the Computer Science and Information Systems course, however six subjects of the BSc sample failed to complete the task and were discarded from the analysis. Table 10 summarizes the experimental set-up.

Table 10: Experimental Design – Experiment 2

Subexperiment	Graph	tabular	Total
UNTIN-MSC	18	17	35
PUCRS-MSC	6	7	13
PUCRS-BSC	12	9	21

A risk model, either tabular or graphical, was randomly distributed between the participants, together with a comprehensibility questionnaire, focused on specific aspects of the risk models.

The experiments were presented as a laboratory activity and only the high level goal of the experiment was highlighted, while the experimental hypotheses were not provided so as not to influence the subjects. The subjects were informed about the experimental procedure.

5.2 Demographics

A preliminary statistical analysis of the sample adopted is shown in Table 11. Remarkably 75% had working experience.

Most of the subjects reported a limited expertise with respect to security knowledge against a good general knowledge of modeling languages. A possible explanation might be that software engineering courses are compulsory in both universities which include lectures on UML and other graphical modeling notations, while security courses are only eligible courses.

Finally subjects had mainly a basic knowledge of the Healthcare application scenario about the HCN.

Table 11: Overall Subjects' Demographic Statistics – Experiment 2

Variable	Scale	Mean/Med.	Distribution
Age	Years	25.8	45% were 19-23 yrs old; 36% were 24-29 yrs old; 19% were 30-46 yrs old
Gender	Sex		78% male; 22% female
Work Experience		3.9	25% had no experience; 43% had 1-3 yrs; 15% had 4-7 yrs; 17% had > 7 yrs
Expertise in Security (median)	0(Novice)-4(Expert)	1	29% novices; 49% beginners; 22% competent users;
Expertise in Modeling Languages (median)		2	13% novices; 22% beginners; 54% competent users; 13% proficient users;
Expertise in HCN (median)		0	67% novices; 23% beginners; 10% competent users

Table 12: Average Precision and Average Recall by subexperiment and risk model type – Experiment 2

	Tabular			Graphical		
	Mean	Median	sd	Mean	Median	sd
Average Precision						
UNTIN-MSC	0.91	0.93	0.06	0.84	0.88	0.12
PUCRS-MSC	0.86	0.92	0.13	0.71	0.74	0.09
PUCRS-BSC	0.86	0.88	0.12	0.79	0.81	0.16
Overall	0.88	0.92	0.10	0.80	0.83	0.14
Average Recall						
UNTIN-MSC	0.89	0.91	0.08	0.77	0.80	0.14
PUCRS-MSC	0.90	0.96	0.12	0.63	0.67	0.13
PUCRS-BSC	0.9	0.92	0.11	0.74	0.79	0.18
Overall	0.90	0.92	0.09	0.74	0.75	0.16

5.3 Results

In this Section the descriptive statistics of the experiment is carried on, firstly analysing Mean, Median and Standard deviation of the metrics already presented in Section 3.2.1, for both the whole group of answers and the single questions. The complexity of the questions is here displayed as well, showing a clear influence on results then further investigated.
The null hypothesis H_0 of equal comprehensibility of the two models is then discarded using aggregated F-measure, making use of the non parametric MannWhitney test which is used to compare two population means that come from the same population.
The correlation between the complexity of the question and the results is finally investigated.

5.3.1 Descriptive Statistics

Table 12 reports the descriptive statistics for results of the application phase of the experiment divided by graphical risk model. It also shows the partial result of each session. Already at this level of analysis the results show that the mean of the average precision of each subject over all the questions is 11% better for the graphical risk model. Similarly, for the average recall of each subject over all the questions, the mean is 23% higher. As the complexity was varied among the questions, precision and recall are also analysed for each question, see Table 13. By comparing the result with the complexity of questions the correlation between it and the mean precision of all the subjects on that question, can be immediately noticed, as the two questions of higher complexity have a much lower precision than the other questions, both in the tabular and graphical model. Interestingly this influence is not evident in the recall, where, the generally lower results registered for the graphical model, actually drop for questions $Q2$, $Q5$, $Q6$, and $Q7$ for no clear reason.

Table 13: Precision and recall by questions and risk model type – Experiment 2

Q#	**Comp-lexity**	**Tabular**			**Graphical**		
		Mean	**Med.**	**sd**	**Mean**	**Med.**	**sd**
Precision							
Q1	2	1.00	1.00	0.00	0.79	1.00	0.37
Q2	4	0.92	1.00	0.25	0.81	1.00	0.40
Q3	2	0.99	1.00	0.06	0.95	1.00	0.19
Q4	2	0.94	1.00	0.24	0.86	1.00	0.35
Q5	6	0.64	1.00	0.46	0.46	0.25	0.48
Q6	2	0.99	1.00	0.06	0.66	1.00	0.44
Q7	4	0.97	1.00	0.10	0.94	1.00	0.20
Q8	4	0.99	1.00	0.06	0.96	1.00	0.18
Q9	2	0.94	1.00	0.24	0.88	1.00	0.32
Q10	4	0.87	1.00	0.27	0.85	1.00	0.31
Q11	4	0.83	1.00	0.29	0.85	1.00	0.31
Q12	6	0.53	0.50	0.27	0.61	0.50	0.35
	Overall	0.88	1.00	0.27	0.80	1.00	0.36
Recall							
Q1	2	0.97	1.00	0.12	0.79	1.00	0.37
Q2	4	0.92	1.00	0.25	0.61	0.50	0.38
Q3	2	1.00	1.00	0.00	0.96	1.00	0.18
Q4	2	0.94	1.00	0.24	0.86	1.00	0.35
Q5	6	0.70	1.00	0.47	0.50	0.50	0.51
Q6	2	0.95	1.00	0.15	0.65	1.00	0.44
Q7	4	0.89	1.00	0.20	0.62	0.75	0.24
Q8	4	0.80	0.67	0.17	0.78	1.00	0.28
Q9	2	0.87	1.00	0.26	0.73	0.80	0.32
Q10	4	0.91	1.00	0.23	0.66	0.67	0.30
Q11	4	0.98	1.00	0.09	0.89	1.00	0.27
Q12	6	0.80	1.00	0.35	0.79	1.00	0.38
	Overall	0.9	1.00	0.25	0.74	1.00	0.36

In questions $Q3$, $Q7$, $Q8$, $Q10$ and $Q11$ the two risk models demonstrated similar precision. For the precision of $Q12$ the graphical risk model slightly outperformed the tabular one. For questions $Q3$, $Q8$ and $Q12$ the two risk models have similar recall. This might be due to the close relationships of questions $Q3$, $Q7$ and $Q8$, or to the easiness of locating of information cues in the diagram like for $Q10$.

5.3.2 Hypothesis Testing

The null hypothesis H_0 of equal comprehensibility of the two models is here discussed using the metrics already defined, and focusing in particular on the aggregated F-measure. A preliminary analysis using the Shapiro-Wilk test shows that our dependent variable, the F-measure, is not normally distributed. Thus, we proceed with non-parametric Mann-Whitney test for the results of the first study (recall that it has been designed as a between-subject test where each examined subject has either a tabular or a graphical model to analyse). This is a non-parametric test that is used to compare two population means that come from the same population.

RQ1: Effect of Risk Model Type on Comprehension

Figure 3 presents the average precision over average recall of the subjects' responses to the comprehension task, divided by Risk Model Type, graphically stressing that for the tabular model the majority of the subjects has an average precision/recall higher than the median values, and the contrary holds for the graphical model. Two finer combination of precision and recall are thus used in Tables 14 and 15. In the former the number of subjects that achieved a comprehension higher (respectively lower) than the median of average F-measure, averaged over both the graphical and tabular model is shown, while in the latter it is the aggregated F-measure that is shown, both confirming the subjects who used tabular risk model achieved better comprehension of the model. In order to validate this result a Fisher's test is thus used, which is a statistical significance test used

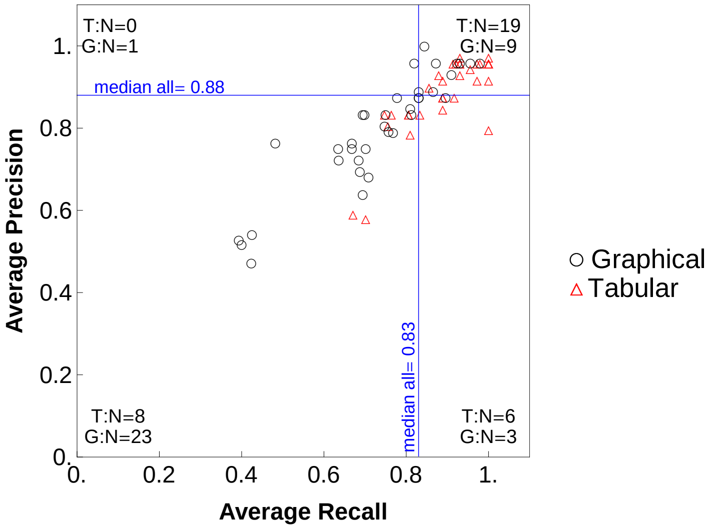


Figure 3: Distribution of Average Precision vs. Average Recall per subject by risk model type – Experiment 2

Table 14: Subjects' Actual Comprehension by Average F-measure by risk model type – Experiment 2

	Tab.	Graph.	Total
$\bar{F}_{m,s} \geq$ median$(\cup_{m,s}\{\bar{F}_{m,s}\})$	23	13	36
$\bar{F}_{m,s} <$ median$(\cup_{m,s}\{\bar{F}_{m,s}\})$	10	23	33
Total	33	36	69

Table 15: Subjects' Actual Comprehension by Aggregated F-measure by risk model type – Experiment 2

	Tab.	Graph.	Total
$tF_{m,s} \geq$ median$(\cup_{m,s}\{tF_{m,s}\})$	23	12	35
$tF_{m,s} <$ median$(\cup_{m,s}\{tF_{m,s}\})$	10	24	34
Total	33	36	69

in the analysis of contingency tables between the two kinds of classification. The results of the test clearly confirms the thesis that the performance of the subjects with respect to actual comprehension is significantly differed by risk model type. For the former (average F-measure) the results of the Fisher's test are $p\text{-}value = 0.008$, confidence interval $CI = [1.34, 12.64]$, and odds ratio is 3.98, while for the latter (aggregated F-measure) it results $p\text{-}value = 0.004$, $CI = [1.5, 14.43]$, and odds ratio 4.49. Where, in particular, the *p-value* represents the probability that the equality thesis H_0 holds true. Very similar results will be shown in the next Sections, for the subexperiment with the students of the Master by Poste Italiane.

Where the reference median median$(\cup_{m,s}\{tF_{m,s}\})$ is evaluated over all the average F-measures of all the subjects, independently from the model received. There is a significant effect of the risk model type on the actual comprehensibility summarized by the average F-measure. Tabular risk model provides better comprehension. The statistical significance of this observation is confirmed by Fisher's tests. Similar results are achieved using the aggregated F-measure as the actual comprehensibility.

Finally it is checked if the rejection of the H_0 thesis holds for each subexperiment. Table 16 reports mean, median, and standard deviation of F-measure by risk model type and subexperiment.

Table 16: Average F-measure by subexperiment and risk model type – Experiment 2

	Tabular			Graphical			
	Mean	**Median**	**sd**	**Mean**	**Median**	**sd**	$\mathbf{Z}_{MW}$
UNTIN-MSC	0.89	0.92	0.07	0.78	0.81	0.13	-2.6 **
PUCRS-MSC	0.87	0.93	0.13	0.65	0.68	0.12	-2.3 *
PUCRS-BSC	0.86	0.89	0.12	0.75	0.79	0.17	-1.6
Overall	0.88	0.90	0.10	0.75	0.77	0.15	-3.9 ***

The last column in Table 16 reports Z statistics and p-value returned by Mann-Whitney test. As we can see from Table 16, in all three subexperiments and overall across them the actual comprehension is higher for the tabular risk model and this is statistically significant for all subexperiments except one (PUCRS-BSC). This means that the subjects who used the tabular risk model achieved overall 17% better actual comprehension than the subjects who used the graphical risk model. The results of Mann-Whitney (MW) test confirmed that the differences in favor of tabular risk model are statistically significant (MW $p\text{-}value = 8.1 * 10^{-5}$; the power of the test is 0.99).

Thus, for what concerns this experiment, the null hypothesis H_0 can be rejected, and we can conclude that tabular risk modeling notation is more effective in supporting comprehension of security risks than the graphical one.

The subjects achieved significantly better comprehension of security risks when used tabular risk model rather than the graphical one. This observation holds across all subexperiments. The significance of the findings is confirmed by a MW test. The level of statistical significance is specified by • ($p < 0.1$), * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

5.3.3 Correlation with question complexity

The relation already shown in Table 13 between the complexity of the questions and the precision/recall, is here further investigated. Questions are divided in simple and complex questions, as they have a complexity $q \leq 2$ or $q > 2$ respectively (see 1 for the definition/calculation of complexity).

For this we need to redefine the metrics above taking into account the complexity of the questions. Thus we will refer to a precision, recall and F-measure of the subject s with risk model type m averaged over all the questions with complexity category q (either $q = 2$ or $q > 2$). The formulation of these is essentially identical to the ones used so far, but for the fact that i only ranges over the question with complexity $q = 2$ or $q > 2$ respectively.

Figure 4 shows the distribution of precision and recall of the subjects made only on the complex questions (see Table 13). There is a significant difference in the recall of the responses to the complex questions of the graphical and tabular models. 76% of the subjects who used tabular risk model achieved better recall than the median value, while only 31% of the subjects for the graphical model passed the median value. The difference in precision is, instead, much smaller.

For completeness, the interaction plots between precision/recall and questions complexity is also shown in Figure 5, to compare the average performances over the simple questions with that over the complex ones. Clearly the complexity of the question deeply influences the precision of the tabular questions, although for both simple and complex questions the tabular risk model has better precision and recall than the graphical one. Tabular risk model has better precision (mean is 0.97) than the graphical one (0.83), while both risk models have similar precision of the responses to the complex questions.

Again, we refine the findings by combining Precision and Recall in the F-measure, averaged over simple and complex questions separately. We count the number of tested subjects which show an average comprehension of the simple/complex questions higher than the global median (evaluated

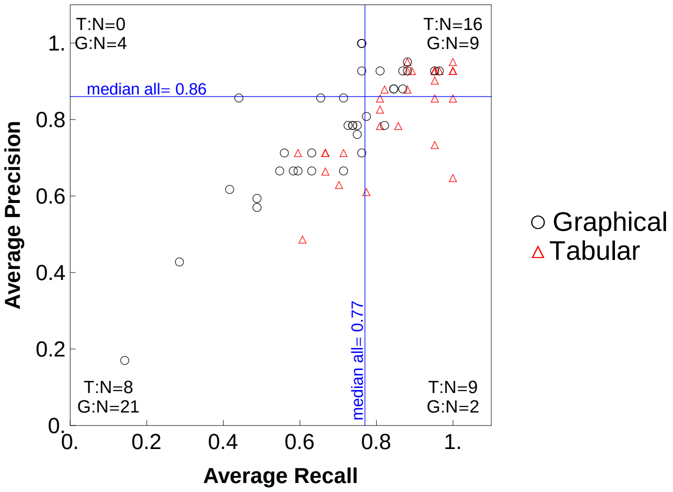


Figure 4: Distribution of Complex Average Precision vs Complex Average Recall per subject by risk model type – Experiment 2

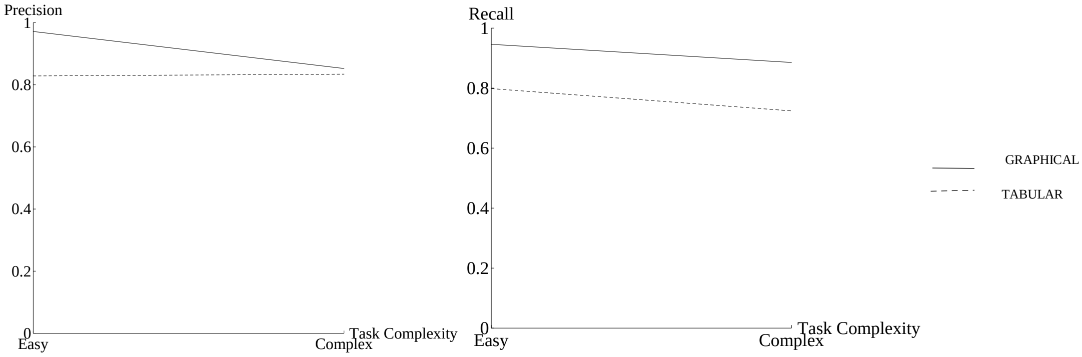


Figure 5: Complexity vs Precision and Recall – Experiment 2

over all the questions). Therefore, Tables 17a, 17b and 17c compare the number of subjects that achieved comprehension on easy/complex higher/lower than the global median, for both the models, for the sole tabular and the sole graphical respectively.

The result found in Fig 5 is here clearly confirmed by Table 17b where it is clear that the easy questions really ease the comprehension of the tabular model. The majority of the subjects (91%) who used tabular risk model showed high comprehension when answering the simple questions and around half of these subjects showed high comprehension of the complex questions. In contrast, only 47% who used graphical risk model showed high comprehension when responding to simple questions and only 19% showed high comprehension of the complex questions. The statistically significant effect of task complexity on the subjects comprehension is also supported by the Fishers test results on the Tables, respectively yielding p-value $= 1.4 \times 10^{-4}$; $CI = [1.86; 8.66]$, OR 3.96, p-value $= 8.1 \times 10^{-4}$; $CI = [2.17; 55.59]$, OR 9.08 and p-value $= 0.02$; $CI = [1, 2; 12, 5]$, OR 3.6.

There is a significant effect of the task complexity on actual comprehensibility summarized by the average F-measure. The majority of the subjects 68% achieved high comprehension when answering simple questions, while only one third of the subjects achieved high comprehension when responding to complex questions. It is even more evident for the subjects who used tabular risk model to

Table 17: Comprehension by question complexity – Experiment 2

(a) Both risk models

	Study		
	Simple Q.	**Complex Q.**	**Total**
$F_{m,s,q} \geq \text{median}(\cup_{m,s}\{F_{m,s,q}\})$	47	24	72
$F_{m,s,q} < \text{median}(\cup_{m,s}\{F_{m,s,q}\})$	22	45	67
Total	69	69	138

(b) Tabular risk model

	Study		
	Simple Q.	**Complex Q.**	**Total**
$F_{m,s,q} \geq \text{median}(\cup_{s}\{F_{m,s,q}\})$	30	17	47
$F_{m,s,q} < \text{median}(\cup_{m,s}\{F_{m,s,q}\})$	3	16	19
Total	33	33	66

(c) Graphical risk model

	Study		
	Simple Q.	**Complex Q.**	**Total**
$F_{m,s,q} \geq \text{median}(\cup_{m,s}\{F_{m,s,q}\})$	17	7	24
$F_{m,s,q} < \text{median}(\cup_{m,s}\{F_{m,s,q}\})$	19	29	48
Total	36	36	72

Table 18: Post-task questionnaire results – Experiment 2

Q#	**Tabular**			**Graphical**		
	Mean	**Med.**	**sd**	**Mean**	**Med.**	**sd**
Q1	0.33	0.00	0.54	0.78	1.00	0.87
Q2	1.12	1.00	1.05	1.14	1.00	1.10
Q3	0.82	1.00	0.68	0.75	1.00	0.65
Q4	1.00	1.00	0.75	0.97	1.00	0.70
Q5	1.00	1.00	0.83	1.31	1.00	0.86
Q6	0.73	1.00	0.76	1.33	1.00	0.89
Q7	0.67	1.00	0.82	0.81	1.00	0.89
Q8	0.70	1.00	0.77	0.47	0.00	0.61

complete the task. This observation is supported by Fishers test.

5.3.4 Post task questionnaire

To control the effect of the experiment settings on the results we analyse subjects feedback collected with post-task questionnaire after the application task. Table 18 presents the descriptive statistics of the responses to post-task questionnaire. Responses are on a 5-item Likert scale from 0 (strongly agree) to 4 (strongly disagree). Overall, for both tabular and graphical risk models subjects concluded that the time allocated to complete the task was enough (Q1). Subjects who used tabular risk model were more confident in adequacy of allocated time than the subjects who used graphical risk model. They found the objectives of the study (Q2) and the task (Q3) clear. In general, the subjects were confident that the comprehension questions are clear (Q4) and they did not experienced difficulty to answer the comprehension questions (Q5). Also both groups did not experienced significant difficulties in understanding (Q6) and using electronic version (Q7) of risk model tables or diagrams. The online survey tool was also easy to use (Q8).

Since we provided subjects with electronic version of tabular and graphical risk models, we decided to investigate whether the subjects used search/filtering information in tables and diagrams. Most of the subjects 64% who used tabular risk models also used search or filtering information in browser or MS Excel, while only half of the subjects who used graphical risk model search in PDF.

6 Experiment 3: comprehensibility of risk models with MSC students

Experiment 3 is somehow close to Experiment 2. The main difference is in the design of the two experiments, indeed, while the previous has a between-subjects design (recall that it means that each examined subject has either a tabular or a graphical model to analyse), this is a within-subjects one. This means that each subject is tested on both a tabular and a graphical representation of the risk model. Of course there is a trade off between the advantages measuring the actual comprehension of the tabular or graphical model on a single subject, and the possible learning effect. In order to mitigate such unwanted effect, this experiment was designed as a with two factors, instead of one, adding to the choice between risk modeling notation, (graphical/tabular) of the Experiment 2 the dual application scenario: either online banking or HCN. To mitigate a possible effect of the treatments' order on the experimental results, the Latin Square was used: in the first session of each experiment the subjects where randomly assigned a tabular or graphical model, while in the second sessions they were assigned the complementary model.

6.1 Experiment execution

Two controlled experiments of two sessions each, both involving MSc students were conducted for this experiment. The former at University of Calabria in Cosenza in September 2015. The subjects were 52 students attending a professional master course in Cybersecurity. The experiment was presented as an entry evaluation activity for the course and only the high level goal of the experiment was discovered. The same setting were kept in the replication conducted at the University of Trento in October 2015 as part of the Security Engineering course. It involved 51 MSc students in Computer Science and was presented as a laboratory activity.

The comparison of two types of risk models was done using questionnaire about comprehensibility of specific aspects of risk models that were distributed to the participants. The subjects were instructed about the experimental procedure. Some subjects who failed to complete both sessions or had a problem with SurveyGizmo platform. As such, 10 subjects from each replication where discarded from the results thus resulting in the experimental setup reported in Table 19.

Table 19: Experimental Design – Experiment 3

Experiment	Graph	tabular	Total people
POSTE	41	41	41
UNITN	42	42	42

6.2 Demographics

Table 20 summarizes the demographic information about the subjects of our experiments. Half of the subjects reported that they had working experience.

The subjects of the study had slightly better security knowledge and slightly worse knowledge of modeling languages comparing to the subjects of the Experiment 2 (see Table 11). They also had basic knowledge of the application scenarios, yet better than the previous Experiment.

6.3 Results

I this Section the data collected for the experiment are here analysed, again with the aim of testing the comprehensibility difference between the tabular and the graphical method. In order to keep the

Table 20: Overall Subjects' Demographic Statistics – Experiment 2

Variable	Scale	Mean/Med.	Distribution
Age	Years	26.4	25% were 21-23 yrs old; 55% were 24-29 yrs old; 20% were 30-40 yrs old
Gender	Sex		75% male; 25% female
English Level	A1 - C2		1% Elementary (A1); 5% Pre-Intermediate (A2); 37% Intermediate (B1); 31% Upper-Intermediate (B2); 15% Advanced (C1); 11% Proficient (C2)
Work Experience		1.3	49% had no experience; 39% had 1-3 yrs; 11% had 4-7 yrs; 1% had > 7 yrs
Expertise in Security (median)	0(Novice)-4(Expert)	1	19% novices; 52% beginners; 18% competent users; 6% proficient; 5% experts
Expertise in Modeling Languages (median)		2	16% novices; 33% beginners; 36% competent users; 13% proficient users; 2% experts
Expertise in Online Banking (median)		0	73% novices; 21% beginners; 4% competent users; 1% proficient users; 1% experts
Expertise in HCN (median)		0	81% novices; 18% beginners; 1% experts

result comparable with that of the previous experiment, the scheme of the analysis performed is the same as the previous. Therefore, for the seek of brevity, the explanation of the motivations of the analysis is kept as short as possible, while the focus of the Section is more on the comparison with the previous results.

6.3.1 Descriptive Statistics

The descriptive statistic for precision and recall of the experiment is shown in Table 21 of application phase of the experiment.

The results of this preliminary analysis do not seem to be affected by the redesign of the test, as again the tabular risk model shows an overall 11% better precision and an overall 25% better recall over the responses given with the graphical risk model, which is very close to the results of Experiment 2. Surprisingly, the results for the recall are, worse than the previous experiment, although the questions are, in general, easier, as shown in Table 21 where precision and recall by questions are listed. However, since the precision remains approximately the same, an analysis of the F-measure is required.

Table 21: Average Precision and Average Recall by Session and risk model type – Experiment 3

	Tabular			Graphical		
	Mean	Median	sd	Mean	Median	sd
Precision						
POSTE	0.91	0.94	0.1	0.79	0.89	0.20
UNITN	0.90	0.92	0.12	0.80	0.81	0.17
Overall	0.90	0.92	0.11	0.80	0.83	0.18
Recall						
POSTE	0.87	0.90	0.12	0.65	0.71	0.22
UNITN	0.87	0.88	0.12	0.70	0.72	0.20
Overall	0.87	0.89	0.12	0.68	0.71	0.21

Remarkably, Table 22 shows that, when measured on the same subjects, the recall for the graphical model spans on a range that is much lower on the majority of the questions. Furthermore it must be considered that the questions in this experiment are in general easier, and that the complexity of the single question does not seem to show an easy to spot correlation with the recall. This result is better investigated in the next subsections.

Table 22: Precision and recall by questions and risk model type– Experiment 3

Q#	Comp-lexity	Tabular			Graphical		
		Mean	Median	sd	Mean	Median	sd
Precision							
Q1	2	0.93	1.00	0.24	0.64	1.00	0.48
Q2	2	0.94	1.00	0.24	0.94	1.00	0.23
Q3	3	0.98	1.00	0.12	0.99	1.00	0.07
Q4	3	0.95	1.00	0.20	0.88	1.00	0.32
Q5	4	0.99	1.00	0.07	0.90	1.00	0.28
Q6	4	1.00	1.00	0.03	0.99	1.00	0.08
Q7	3	0.87	1.00	0.34	0.73	1.00	0.44
Q8	3	0.95	1.00	0.18	0.71	1.00	0.44
Q9	4	0.84	1.00	0.30	0.88	1.00	0.24
Q10	4	0.64	1.00	0.48	0.42	0.00	0.50
Q11	5	0.93	1.00	0.19	0.84	1.00	0.32
Q12	5	0.84	1.00	0.37	0.64	1.00	0.48
	Overall	0.90	1.00	0.28	0.80	1.00	0.39
Recall							
Q1	2	0.94	1.00	0.24	0.64	1.00	0.48
Q2	2	0.93	1.00	0.25	0.76	1.00	0.29
Q3	3	0.99	1.00	0.11	0.96	1.00	0.14
Q4	3	0.87	1.00	0.25	0.62	0.67	0.30
Q5	4	0.94	1.00	0.15	0.64	0.75	0.32
Q6	4	0.86	1.00	0.17	0.60	0.60	0.20
Q7	3	0.87	1.00	0.34	0.73	1.00	0.44
Q8	3	0.96	1.00	0.18	0.64	0.67	0.42
Q9	4	0.77	1.00	0.33	0.81	1.00	0.29
Q10	4	0.64	1.00	0.48	0.42	0.00	0.50
Q11	5	0.84	1.00	0.25	0.67	0.50	0.32
Q12	5	0.84	1.00	0.37	0.64	1.00	0.48
	Overall	0.87	1.00	0.29	0.68	1.00	0.39

6.3.2 Hypothesis Testing

As in the Experiment 2 the necessary preliminary Shapiro-Wilk test on the F-measure reveals that its distribution is not Normal, thus choosing the type of analysis to be performed on it. Figure 6 presents, for the comprehension task, the average precision and recall of the responses while Table 23 compares the number of subjects that achieved better comprehension than the median average F-measure with the number of subjects that have lower comprehension level. With respect to Figure 6, the drop in recall for the graphical model is here made evident by the relevant presence of black circles shifted on the left side of the graph. The better performance of the tabular model is again easy to spot.

This result is confirmed in Table 24 and the relative Fisher's test on its findings: the performance of the subjects in both studies with respect to actual comprehension is significantly affected by risk model type ($p\text{-}value = 4.9 * 10^{-7}, CI = [2.64, 11.04]$, odds ratio 5.32). We made the robustness check with aggregated F-measure, Table 24 with its relative Fisher's test ($p\text{-}value = 5.1 * 10^{-9}, CI = [3.46, 15.21]$, the odds ratio is 7.13).

There is a significant effect of the risk model type on the actual comprehensibility summarized by the average F-measure $F_{m,s}$ for model m and subject s. Tabular risk model provides better comprehension. The statistical significance of this observation is confirmed by Fisher's tests. Similar results are achieved using the aggregated F-measure as the actual comprehensibility.

We also checked if this finding holds for each individual experiment. Table 25 measure by risk model type and experiment for two studies. The last two columns in Table 25 report Z statistics and p-value returned by Wilcoxon test. The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test; it is the equivalent of the Mann-Whitney test for the case in which the samples are related, or even matched, or the measurements are repeated on the same sample to assess whether their population mean ranks differ (i.e. it is a paired difference test).

Once again the results shown are very close to those obtained for the Experiment 2 (see Table 14).

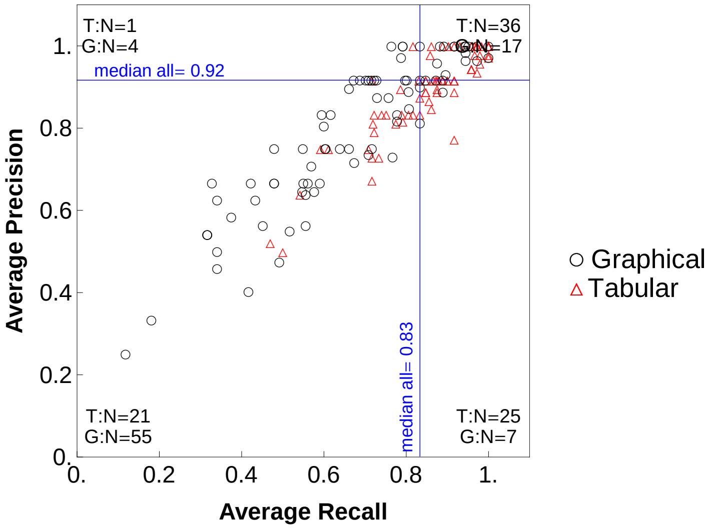


Figure 6: Distribution of Average Precision vs Average Recall per subject by risk model type – Experiment 3

Table 23: Subjects' Actual Comprehension by Average F-measure by risk model type – Experiment 3

	Tab.	**Graph.**	**Total**
$F_{m,s} \geq \text{median}(\cup_m \{F_{m,s}\})$	58	25	83
$F_{m,s} < \text{median}(\cup_m \{F_{m,s}\})$	25	58	83
Total	83	83	166

Table 24: Subjects' Actual Comprehension by Aggregated F-measure by risk model type – Experiment 3

	Tab.	**Graph.**	**Total**
$tF_{m,s} \geq \text{median}(\cup_m \{tF_{m,s}\})$	61	23	84
$tF_{m,s} < \text{median}(\cup_m \{tF_{m,s}\})$	22	60	82
Total	83	83	166

Table 25: Average F-measure by subexperiment and risk model type – Experiment 3

	Tabular			**Graphical**			
	Mean	**Median**	**sd**	**Mean**	**Median**	**sd**	$\mathbf{Z}_W$
POSTE	0.88	0.92	0.11	0.69	0.76	0.21	-4.77 ***
UNITN	0.88	0.90	0.12	0.73	0.75	0.19	-3.54 ***
Overall	0.88	0.90	0.12	0.71	0.76	0.20	-5.98 ***

The subjects who used the tabular risk model achieved overall 18% better actual comprehension than the subjects who used the graphical risk model. The results of Wilcoxon test confirmed that the differences in favor of tabular risk model are statistically significant (Wilcoxon *p-value* $= 6.3 * 10^{-10}$; the power of the test is 1).

Thus, also in the within subject designed Experiment the null hypothesis H_0 can be rejected, and we can conclude that tabular risk modeling notation is more effective in supporting comprehension of security risks than the graphical one.

The subjects achieved significantly better comprehension of security risks when used tabular risk model rather than the graphical one. This observation holds across all experiments of two studies.

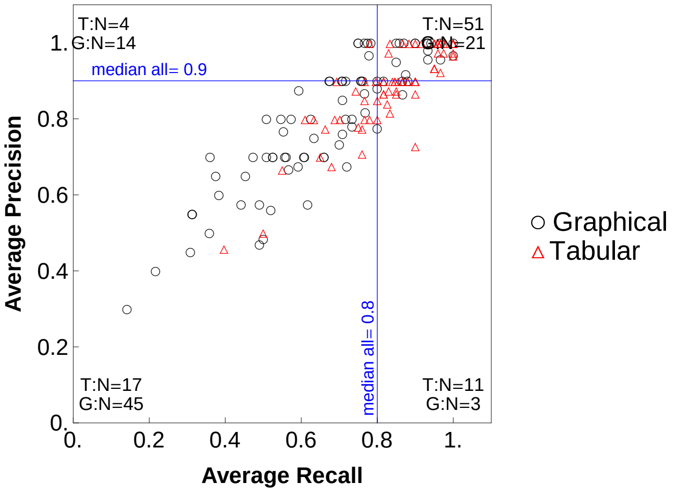


Figure 7: Distribution of Complex Average Precision vs. Complex Average Recall per subject by risk model type – Experiment 3

The significance of the findings is confirmed by a Wilcoxon test in the second study. The level of statistical significance is specified by • ($p < 0.1$), * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

6.3.3 Correlation with question complexity

Figure 7 shows the distribution of precision and recall of the subjects made only on the complex questions, i.e. the questions with complexity > 2 (see 22). There is a significant difference in the recall of the responses to the complex questions of the graphical and tabular models. 76% of the subjects who used tabular risk model achieved better recall than the median value, while only 31% of the subjects for the graphical model passed the median value. The difference in precision is, instead, much smaller.

We show the interaction plots between precision, recall, and individual questions complexity. Figure 8 shows the significant interaction between risk model type, questions complexity and recall in favor of the tabular model for both simple and complex questions. About the difference in precision, in contrast to the results of the previous experiment, there is no drop in comprehension of the complex question for the tabular model. Moreover, for the simple questions tabular risk model has better precision (mean is 0.93) than the graphical one (0.79), while for the complex questions the mean precision is 0.9 and 0.8 respectively for tabular and graphical risk models.

To make this analysis more precise we calculate the F-measure by questions complexity. So that $F_{m;s;q}$ is the mean value for the subject s using risk model m over all questions i with complexity level which can be either $q \leq 2$ or $q > 2$. Table 26a compares the number of subjects that achieved high comprehension with the ones that achieved a low comprehension.

The majority of the subjects (84%) that used tabular risk model showed high comprehension when answering the simple questions but this time the percentage of subjects that showed high comprehension of the complex questions is higher than in Experiment 2. This effect is statistically significant (Fishers test $p-value = 4.3 \times 10^{-3}$, OR 3.03). In contrast, results for the graphical model are pretty close in percentage to the ones of the previous experiment, thus lower than the tabular in general. Again Fishers tests confirm the statistical significance of the result found.

There is a significant effect of the task complexity on actual comprehensibility summarized by the average F-measure. The majority of the subjects 68% achieved high comprehension when answering

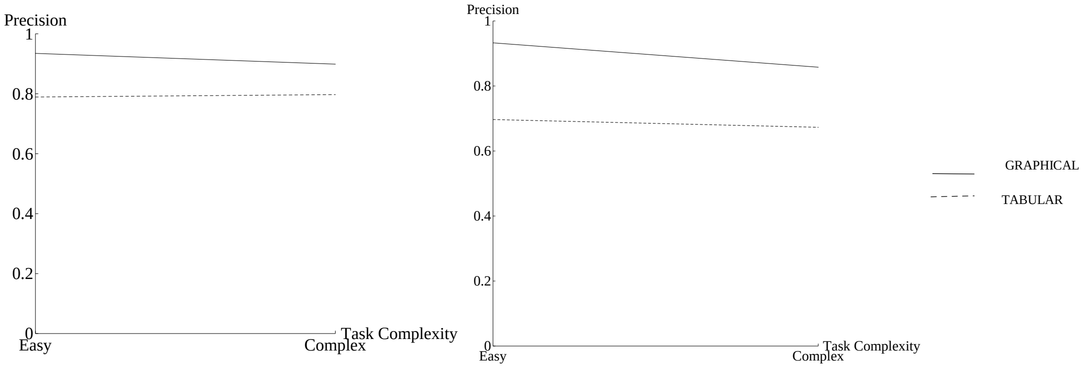


Figure 8: Complexity vs Precision and Recall – Experiment 3

Table 26: Comprehension by question complexity – Experiment 3

(a) Both risk models

	Study		
	Simple Q.	**Complex Q.**	**Total**
$F_{s,q} \geq \text{median}(\cup_{q,s}\{F_{s,q}\})$	114	72	186
$F_{s,q} < \text{median}(\cup_{q,s}\{F_{s,q}\})$	52	94	146
Total	166	166	332

(b) Tabular risk model

	Study		
	Simple Q.	**Complex Q.**	**Total**
$F_{m,s,q} \geq \text{median}(\cup_{q,s}\{F_{s,q}\})$	70	53	123
$F_{m,s,q} < \text{median}(\cup_{q,s}\{F_{s,q}\})$	13	30	43
Total	83	83	166

(c) Graphical risk model

	Study		
	Simple Q.	**Complex Q.**	**Total**
$F_{m,s,q} \geq \text{median}(\cup_{q,s}\{F_{s,q}\})$	44	19	63
$F_{m,s,q} < \text{median}(\cup_{q,s}\{F_{s,q}\})$	39	64	103
Total	83	83	166

simple questions, while only one third of the subjects achieved high comprehension when responding to complex questions. It is even more evident for the subjects who used tabular risk model to complete the task. This observation is supported by Fishers test. Tabular risk model shows higher average precision and average recall in each experiment and in general, among all the participants.

6.3.4 Post-task questionnaire

To control the effect of the experiment settings on the results we analyse subjects feedback collected with post-task questionnaire after the application task. Table 27 present descriptive statistics of the responses to post-task questionnaire. Responses are on a 5-item Likert scale from 0 (strongly agree) to 4 (strongly disagree). Overall, for both tabular and graphical risk models subjects concluded that the time allocated to complete the task was enough (Q1). Subjects who used tabular risk model were more confident in adequacy of allocated time than the subjects who used graphical risk model. They found the objectives of the study (Q2) and the task (Q3) clear. In general, the subjects were confident that the comprehension questions are clear (Q4) and they did not experienced difficulty

Table 27: Post-task questionnaire results – Experiment 3

Q#	Tabular			Graphical		
	Mean	Med.	sd	Mean	Med.	sd
Q1	0.78	1.00	0.83	1.13	1.00	1.01
Q2	1.14	1.00	0.84	1.29	1.00	0.94
Q3	0.90	1.00	0.77	1.23	1.00	0.93
Q4	1.07	1.00	0.82	1.53	1.00	0.94
Q5	1.08	1.00	0.80	1.43	1.00	0.95
Q6	1.02	1.00	0.78	1.48	1.00	1.11
Q7	0.98	1.00	0.87	1.25	1.00	0.99
Q8	0.96	1.00	0.77	1.14	0.00	0.93

to answer the comprehension questions (Q5). Also both groups did not experienced significant difficulties in understanding (Q6) and using electronic version (Q7) of risk model tables or diagrams. The online survey tool was also easy to use (Q8). Since we provided subjects with electronic version of tabular and graphical risk models, we decided to investigate whether the subjects used search/filtering information in tables and diagrams. almost half of the subjects 45% who used tabular risk models and 39% of the subject who used graphical risk models also used search or filtering information in browser or MS Excel.

7 Experiment 4: comprehensibility of risk models with ATM professionals

Experiment 4 differs from the previous as it is targeting the professionals of the ATM domain. The difficulty in gathering a huge number of professionals in the ATM domain made the size of the sample small, but yet good enough to be compared with the previous Experiments and drive some conclusions. Analogously with Experiment 2, this Experiment is designed as a *between-subject* experiment, where each examined subject has either a tabular or a graphical model to analyse. The test is a single factor (risk modeling notation), double treatment (graphical/tabular). The questions for the test, this time where only 8.

Results are analysed also taking into account the Complexity level of the questions evaluated by the number of logical connections, clues and judgements contained in the questions.

Some basic analysis of the demographical features and the answers to the post-task questionnaire of the sample examined are first taken into account, in order to compare the samples with the other experiments and stress possible influences on the result.

7.1 Experiment execution

The experiment was organised during the SESAR Innovation Days (SID) 2015 in Bologna to attract ATM and Security Researchers and Professionals. SID Participants were invited to an interactive training workshop about Security Risk Assessment Methodologies in ATM. Firstly, a training session about the SecRAM and CORAS Security Risk Assessment Methodologies applied to relevant security scenarios was delivered by SESAR WP16.06.02 members and by Trento University, respectively. Secondly, participants were involved in guided hands-on sessions about comprehensibility of tabular and graphical risk models.

The material collected during the hands-on sessions was the basis of our analysis. The experiment involved a total of 14 ATM professionals with different backgrounds working in both academy and industry. The participants were randomly divided in two groups composed by 7 professionals in each (Group A and B, respectively). Each participant worked individually. The participants belonging to Group A worked with tabular risk models, while the participants in Group B worked with graphical risk models. Each group had to apply the method on the same scenario, namely the On-line Banking Scenario. Table 28 shows the setup of the experiment.

Table 28: Experimental Design – Experiment 4

Experiment	Graph	Tabular	Total people
SID	7	7	14

7.2 Demographics

Table 29 summarizes the demographic information about the subjects of our experiments. All subjects reported that they had working experience with a mean of 18 years. Half of the examined professionals had a competent or higher level for in security, modeling languages and in the Online banking domain, thus reflecting the aim of the Experiment to test highly skilled people of the field.

Table 29: Overall Subjects' Demographic Statistics – Experiment 4

Variable	Scale	Mean/Med.	Distribution
Age	Years	43.2	7% were 24-29 yrs old; 43% were 30-40 yrs old; 50% were >46 yrs old
Gender	Sex		86% male; 14% female
Work Experience		18	43% had 4-7 yrs; 57% had > 7 yrs
Expertise in Security (median)	0(Novice)-4(Expert)	2	29% beginners; 43% competent users; 14% proficient; 8% experts
Expertise in Modeling Languages (median)		2	43% beginners; 36% competent users; 14% proficient users; 7% experts
Expertise in Online Banking (median)		2	7% novices; 14% beginners; 43% competent users; 14% proficient users; 22% experts

7.3 Results

7.3.1 Descriptive Statistics

Both the models show a poor precision and recall. The first result to be noticed in Table 30 is that both the models show a poor precision and recall although the graphical risk model was slightly easier to comprehend than the tabular. Such a small difference must however be analysed by recalling that the sample was a very small one, and must therefore be taken with a pinch of salt.

Table 30: Average Precision and Average Recall by risk model type – Experiment 4

	Tabular			Graphical		
	Mean	Median	sd	Mean	Median	sd
Precision	0.61	0.67	0.28	0.76	0.88	0.3
Recall	0.62	0.71	0.27	0.72	0.88	0.32
F-measure	0.6	0.68	0.28	0.73	0.83	0.32

The general drop in the results for both Precision and Recall, is even more evident looking at the single questions (Table 31). In the case of $Q3$ and $Q6$ the graphical model performs relevantly better than the tabular, with no apparent reason.

7.3.2 Hypothesis Testing

As Precision and Recall assume always approximately the same value, the Markers of Figure 9 appear almost aligned on the diagonal.

The Average (Table 32) and Aggregated (Table 33) F-measure show an identical result, but this time the results of the Fisher's test is $p = 1$, which might be due to the fact that the sample is small.

The result shown in the third line of Table 30 combines together the precision and the recall of the experiment, giving evidence to the fact that in this experiment both the models performed worse than all the previous experiment. The drop affects particularly the tabular model.

7.3.3 Correlation with question complexity

Figures 10 and 11, show that, in this Experiment the complexity does not influence the performance of the models, which actually perform better for Complex questions. This might be due to the fact that we compute complexity as the number of logical connectors contained into a question. This might affect the non professionals more than the professionals, as the novices of the domain stick to the the question, and its logical links, while the professionals go beyond the words contained in the question and apply a different logic path to find the answer.

We report the Fisher's test of Tables 34a to 34c although their high $p - values$ might suggest that true odds ratio might not be 1:

Table 31: Precision and recall by questions and risk model type– Experiment 4

Q#	Comp-lexity	Tabular			Graphical		
		Mean	Median	sd	Mean	Median	sd
Precision							
Q1	2	0.57	1.00	0.53	0.57	1.00	0.53
Q2	5	0.7	0.6	0.29	0.79	1.00	0.27
Q3	3	0.29	0.00	0.49	0.71	1.00	0.49
Q4	5	0.57	1.00	0.53	0.86	1.00	0.38
Q5	4	0.86	1.00	0.38	0.86	1.00	0.38
Q6	4	0.36	0.00	0.48	0.57	0.5	0.45
Q7	4	0.79	1.00	0.39	0.86	1.00	0.38
Q8	2	0.71	1.00	0.49	0.86	1.00	0.38
Recall							
Q1	2	0.57	1.00	0.53	0.57	1.00	0.53
Q2	5	0.9	1.00	0.16	0.86	1.00	0.26
Q3	3	0.29	0.00	0.49	0.71	1.00	0.49
Q4	5	0.57	1.00	0.53	0.86	1.00	0.38
Q5	4	0.71	0.67	0.36	0.57	0.67	0.32
Q6	4	0.43	0.00	0.53	0.71	1.00	0.49
Q7	4	0.81	1.00	0.38	0.75	1.00	0.43
Q8	2	0.64	1.00	0.48	0.76	1.00	0.37

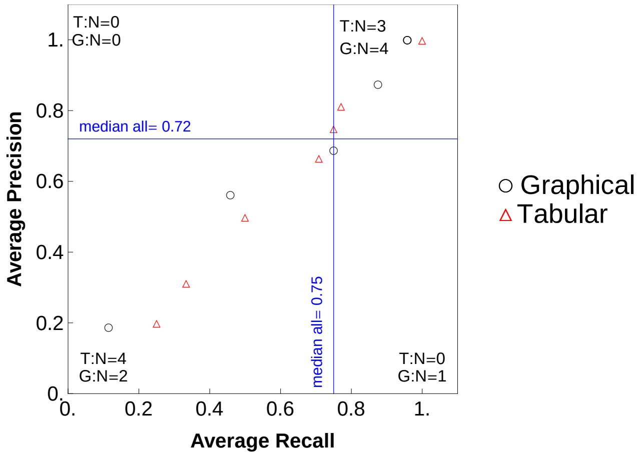


Figure 9: Distribution of Average Precision vs Average Recall per subject by risk model type – Experiment 4

Table 32: Subjects' Actual Comprehension by Average F-measure by risk model type – Experiment 4

	Study		
	Tab.	Graph.	Total
$F_{m,s} \geq \text{median}(\cup_m\{F_{m,s}\})$	3	4	7
$F_{m,s} < \text{median}(\cup_m\{F_{m,s}\})$	4	3	7
Total	7	7	14

Table 33: Subjects' Actual Comprehension by Aggregated F-measure by risk model type – Experiment 4

	Study		
	Tab.	Graph.	Total
$tF_{m,s} \geq \text{median}(\cup_m\{tF_{m,s}\})$	3	4	7
$tF_{m,s} < \text{median}(\cup_m\{tF_{m,s}\})$	4	3	7
Total	7	7	14

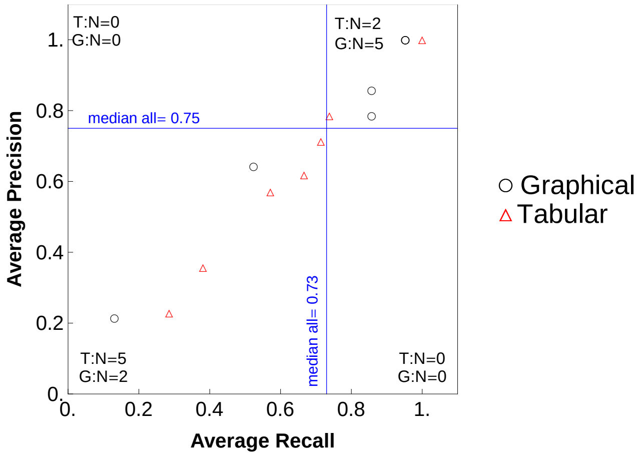


Figure 10: Distribution of Complex Average Precision vs. Complex Average Recall per subject by risk model type – Experiment 4

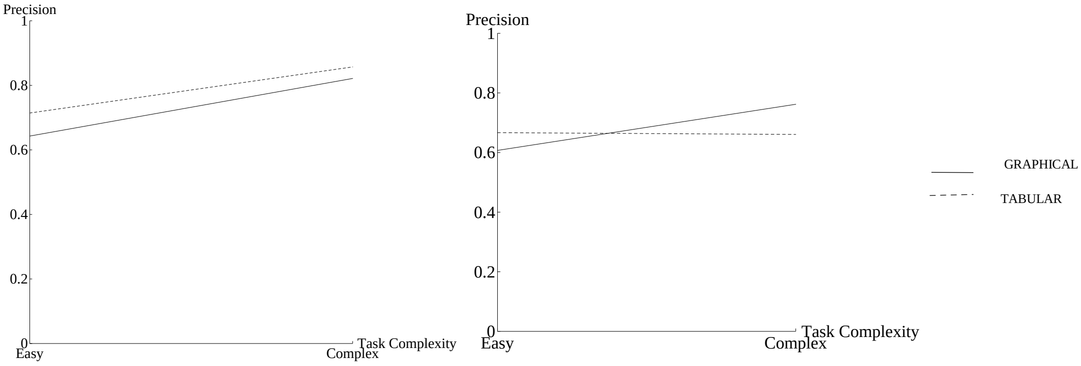


Figure 11: Complexity vs Precision and Recall – Experiment 4

$p - value = 1$, $CI = [0.17313085.5072453]$, $OR = 0.9803083$
$p - value = 0.5921$, $CI = [0.241880155.2697930]$, $OR = 3.043639$
$p - value = 1$, $CI = [0.031137197.71793303]$, $OR = 0.5581105$.

On simple questions the tabular and graphical models perform the same, the difference is in the complex ones.

7.3.4 Post-task questionnaire

Although the performance of both models is worse than all the previous experiments, the post-task questionnaire did not show any particular issue or difficulty in understanding the task or the questions. This is shown in Table 35. Notice that for this experiment we did not ask about the difficulty in using SurveyGizmo and whether or not they used the search function in the PDF document.

Table 34: Comprehension by question complexity – Experiment 4

(a) Both risk models

	Study		
	Simple Q.	**Complex Q.**	**Total**
$F_{s,q} \geq \text{median}(\cup_{q,s}\{F_{s,q}\})$	8	7	15
$F_{s,q} < \text{median}(\cup_{q,s}\{F_{s,q}\})$	6	7	13
Total	14	14	28

(b) Tabular risk model

	Study		
	Simple Q.	**Complex Q.**	**Total**
$F_{m,s,q} \geq \text{median}(\cup_{q,s}\{F_{s,q}\})$	4	2	6
$F_{m,s,q} < \text{median}(\cup_{q,s}\{F_{s,q}\})$	3	5	8
Total	7	7	14

(c) Graphical risk model

	Study		
	Simple Q.	**Complex Q.**	**Total**
$F_{m,s,q} \geq \text{median}(\cup_{q,s}\{F_{s,q}\})$	4	5	9
$F_{m,s,q} < \text{median}(\cup_{q,s}\{F_{s,q}\})$	3	2	5
Total	7	7	14

Table 35: Post-task questionnaire results – Experiment 4

Q#	**Tabular**			**Graphical**		
	Mean	**Med.**	**sd**	**Mean**	**Med.**	**sd**
Q1	0.71	1.00	0.76	0.43	0.00	0.535
Q2	1.29	1.00	1.38	1.14	1.00	0.9 0
Q3	1.71	2.00	1.25	1	1.00	0.58
Q4	1.57	1.00	1.13	1.71	2.00	0.756
Q5	2.43	2.00	1.13	1.29	1.00	0.95
Q6	1.43	2.00	0.787	1.00	1.00	0.577
Q7	1.29	1.00	1.5	0.71	1.00	0.48

8 Experiment 5: comprehensibility of risk models – online experiment with professionals

Experiment 5 was released online, opened to all the people willing to get tested. The invitation for the experiment, however, was only sent to experts of the security domain.
The main difference with previous Experiments is in the introduction of a second graphical model, more precisely an UML model, where each category is also characterized by a symbol. As such this experiment is a *between-subject* experiment, where each examined subject has either a tabular or a graphical or an UML model to analyse. The experiment is a single factor (risk modeling notation), triple treatment (tabular/graphical/UML). Results are analysed also taking into account the complexity level of the questions evaluated by the number of logical connections, clues and judgments contained in the questions.
Some basic analysis of the demographical features and the answers to the post-task questionnaire of the sample examined are first taken into account, in order to compare the samples with the other experiments and stress possible influences on the result.

8.1 Experiment execution

The experiment was opened and accessible online with no requirements. An invitation to submit the questionnaire was sent to experts of the Security field in general, but also some students registered are present in the sample. The online experiment participants were randomly assigned to the one of three modeling notation by the computer. Then, they were asked to complete a demographics and background questionnaire and see a video explaining the modeling notation and the application scenario, namely the On-line Banking Scenario. After, the participants were asked to complete the comprehension task using the corresponding model of Online Banking security risks, and finally the usual post-task questionnaire to be completed.

Some subjects did not finish the task and thus we discarded their results. In total 56 subjects completed the task. Section 8.1 shows the setup of the experiment, for brevity we refer to the graphical model designed by UML as the UML model. The material collected during the sessions was the basis of our analysis. Each subject worked individually.

Table 36: Experimental Design – Experiment 5

Experiment	Graph	UML	tabular	Total
ONLINE	18	19	19	56

8.2 Demographics

Table 37 summarizes the demographic information about the subjects of our experiments. Almost all subjects (97%) reported that they had working experience.

As expected Table 37 shows a general lower level of expertise than in the Experiment 4, as the most of the participants define their level as novice.

Table 37: Overall Subjects' Demographic Statistics – Experiment 5

Variable	Scale	Mean/Med.	Distribution
Age	Years	35.1	27% were 24-29 yrs old; 48% were 30-39 yrs old; 14% were 40-49 yrs old; 11% were 50-62 yrs old
Gender	Sex		75% male; 25% female
English Level	A1 - C2		12.5% Intermediate (B1); 20% Upper-Intermediate (B2); 30% Advanced (C1); 37.5% Proficient (C2) or native speaker
Work Experience		9.7	0.2% had no experience; 29% had 1-4 yrs; 39% had 5-10 yrs; 21% had 11-20 yrs; 11% had > 20 yrs
Expertise in Security (median)	0(Novice)-4(Expert)	1	43% novice; 23% beginners; 23% competent users; 9% proficient; 2% experts
Expertise in Modeling Languages (median)	0(Novice)-4(Expert)	2	21% novice; 27% beginners; 27% competent users; 16% proficient users; 9% experts
Expertise in Online Banking (median)	0(Novice)-4(Expert)	2	20% novices; 23% beginners; 23% competent users; 27% proficient users; 7% experts

8.3 Results

Due to the technical problems in the setup of comprehension task in SurveyGizmo platform for the graphical and UML risk models the responses to the questions Q4, Q6, Q8, Q9, and Q11 were affected. Therefore, we excluded the responses to these questions for all three treatments. Further, the analysis is based only on the responses to the questions Q1-Q3, Q5, Q7, Q10, Q12.

8.3.1 Descriptive Statistics

Table 38 shows the preliminary analysis of the experiment. Although the UML performs slightly better than the graphical model, results of the Experiments 2 and 3 are confirmed: the tabular model shows a better precision and recall than both the graphical methods.

Table 38: Average Precision and Average Recall – Experiment 5

	Tabular			Graphical			UML		
	Mean	Median	sd	Mean	Median	sd	Mean	Median	sd
Precision	0.94	1.00	0.11	0.64	0.64	0.27	0.79	0.86	0.21
Recall	0.94	1.00	0.11	0.60	0.64	0.29	0.77	0.86	0.23

Table 39: Precision and Recall by Questions – Experiment 5

Q#	Complexity	Tabular			Graphical			UML		
		Mean	Median	sd	Mean	Median	sd	Mean	Median	sd
Precision										
Q1	2	1.00	1.00	0.00	0.50	0.5	0.51	0.79	1.00	0.42
Q2	2	0.97	1.00	0.11	0.94	1.00	0.24	0.89	1.00	0.27
Q3	3	0.98	1.00	0.08	0.83	1.00	0.38	0.96	1.00	0.15
Q5	4	0.95	1.00	0.23	0.78	1.00	0.43	0.80	1.00	0.38
Q7	3	0.92	1.00	0.25	0.61	1.00	0.50	0.74	1.00	0.45
Q10	4	0.79	1.00	0.42	0.39	0.00	0.50	0.68	1.00	0.45
Q12	5	0.95	1.00	0.23	0.44	0.00	0.51	0.63	1.00	0.50
	Overall	0.95	1.00	0.23	0.64	1.00	0.48	0.79	1.00	0.40
Recall										
Q1	2	1.00	1.00	0.00	0.50	0.50	0.51	0.79	1.00	0.42
Q2	2	1.00	1.00	0.00	0.83	1.00	0.30	0.84	1.00	0.29
Q3	3	1.00	1.00	0.00	0.78	1.00	0.39	0.92	1.00	0.19
Q5	4	0.89	1.00	0.27	0.64	0.75	0.42	0.70	1.00	0.40
Q7	3	0.95	1.00	0.23	0.61	1.00	0.50	0.74	1.00	0.45
Q10	4	0.79	1.00	0.42	0.39	0.00	0.5	0.74	1.00	0.45
Q12	5	0.95	1.00	0.23	0.44	0.00	0.51	0.63	1.00	0.50
	Overall	0.95	1.00	0.23	0.60	1.00	0.47	0.77	1.00	0.40

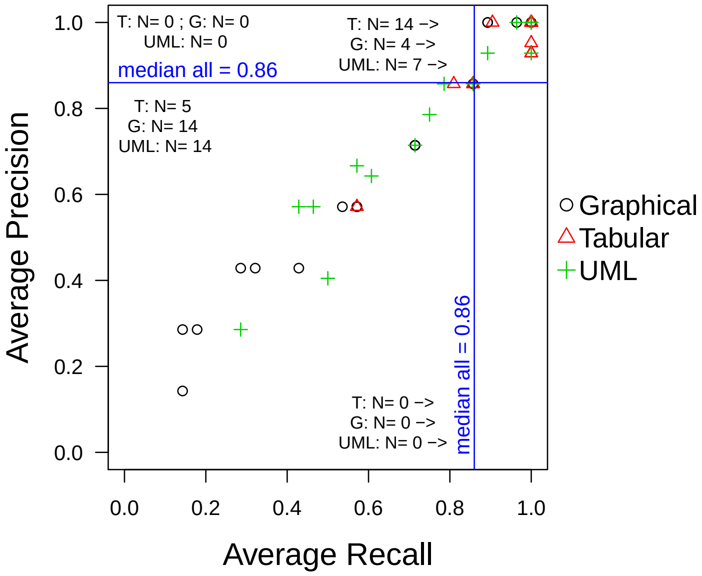


Figure 12: Distribution of Average Precision and Average Recall per Subject by Risk Model Type and Study – Experiment 5

Table 40: Subjects' Actual Comprehension by Average F-measure – Experiment 5

	Tab.	Graph.	UML	Total
$F_{m,s} \geq \text{median}(\cup_m\{F_{m,s}\})$	17	6	10	33
$F_{m,s} < \text{median}(\cup_m\{F_{m,s}\})$	2	12	9	23
Total	19	18	19	56

Table 41: Subjects' Actual Comprehension by Aggregated F-measure – Experiment 5

	Tab.	Graph.	UML	Total
$tF_{m,s} \geq \text{median}(\cup_m\{tF_{m,s}\})$	16	5	8	29
$tF_{m,s} < \text{median}(\cup_m\{tF_{m,s}\})$	3	13	11	27
Total	19	18	19	56

For this experiment the tabular risk model (see Table 39), had a precision and recall both higher than 79% for all the questions, which is the higher result of all the experiments so far. Notice that the UML performed better than the graphical model.

8.3.2 Hypothesis Testing

Figure 12 presents the average precision and recall of the subjects' responses to the comprehension task in the Experiment 5. As we can see, the subjects who used tabular model demonstrated higher comprehensibility than the subjects who used graphical or UML models. Both Average (Table 40) and Aggregated (Table 41) F-measure confirm the very high performance of the tabular model, with their respective results of the Fisher's tests yielding ($p - value = 0.001286$ and $p - value = 0.00127$).

Unlike to the previous experiment, here we compare three different treatments, i.e. we need to compare three unmatched groups. As suggested by Table 5, we can use ANOVA or Kruskal-Wallis (KW) test. However, the results of Shapiro-Wilk test ($p - value = 1.6 * 10^{-6}$) showed that our data

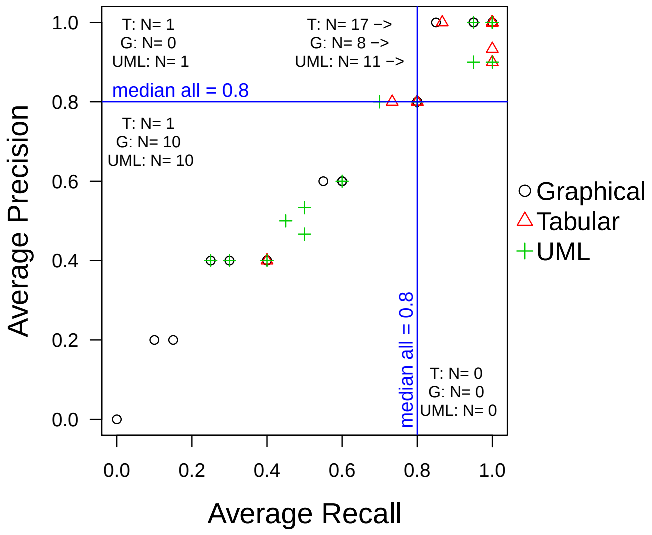


Figure 13: Distribution of Average Precision and Average Recall per Subject by Risk Model Type for Complex Questions – Experiment 5

are not normally distributed, and we have to use KW test.

Table 42 report mean, median, and standard deviation of average F-measure by risk model type. The last column report χ^2 statistic returned by KW test and the level of statistical significance of the results. Table 42 clearly defines the tabular model as the easier to be comprehended for the Experiment 5, and stresses the difference between the graphical and UML models. This is confirmed by KW test ($p - value = 1.36 * 10^{-3}$). To investigate the difference between pairs of treatments we run the post-hoc test using Mann-Whitney test with Holm correction [9, Chap. 14.2]. The MW test revealed statistically difference between the tabular and graphical models ($p - value = 3.3 * 10^{-3}$) and between the tabular and UML models ($p - value = 0.026$), while the difference between the graphical and UML models is not statistically significant ($p - value = 0.24$). Therefore, we can reject the null hypothesis H_0.

Table 42: Average F-measure – Experiment 5

	Tabular			Graphical			UML			KW χ^2
	Mean	Med.	sd	Mean	Med.	sd	Mean	Med.	sd	
Average F-measure	0.94	1	0.11	0.61	0.64	0.28	0.77	0.86	0.22	16.7 ***

8.3.3 Correlation with question complexity

In Figure 13 almost all the red triangles of the tabular model are in the top right quadrant, meaning that on the complex questions almost all the people that showed to comprehend the tabular model better than the median value. The mirror result is instead taken by the graphical model. Indeed, as seen in Figure 14 the performances of the three models are almost equally affected by the complexity of the questions (see Table 39).

The F-measure averaged on the complex questions confirms the findings. Tables 43a to 43d show that the comprehension of the tabular model is very high also on the complex questions. Moreover,

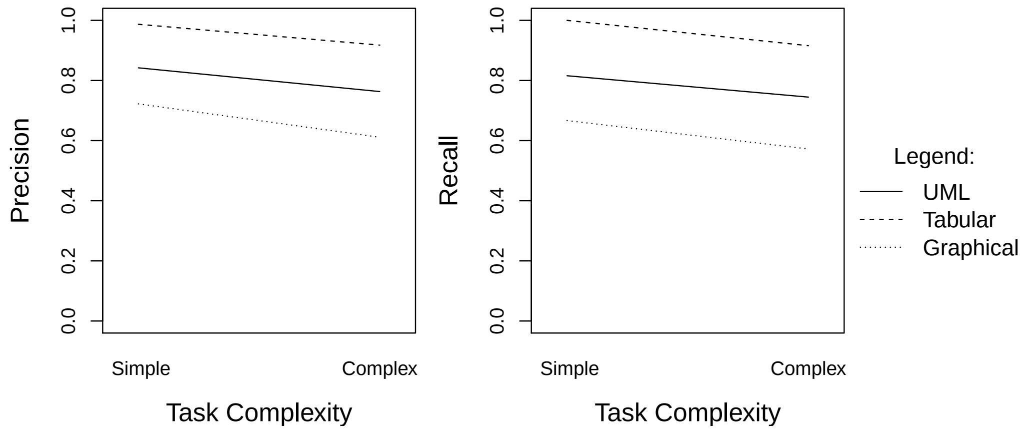


Figure 14: Complexity vs. Precision and Recall – Experiment 5

Table 43: Comprehension by Questions' Complexity – Experiment 5

(a) All risk models

	Simple Q.	**Complex Q.**	**Total**
$F_{s,q} \geq \text{median}(\cup_{q,s}\{F_{s,q}\})$	37	27	64
$F_{s,q} < \text{median}(\cup_{q,s}\{F_{s,q}\})$	19	29	48
Total	56	56	112

(b) Tabular risk model

	Simple Q.	**Complex Q.**	**Total**
$F_{m,s,q} \geq \text{median}(\cup_{q,s}\{F_{s,q}\})$	18	14	32
$F_{m,s,q} < \text{median}(\cup_{q,s}\{F_{s,q}\})$	1	5	6
Total	19	19	38

(c) Graphical risk model

	Simple Q.	**Complex Q.**	**Total**
$F_{m,s,q} \geq \text{median}(\cup_{q,s}\{F_{s,q}\})$	7	4	11
$F_{m,s,q} < \text{median}(\cup_{q,s}\{F_{s,q}\})$	11	14	25
Total	18	18	36

(d) UML risk model

	Study		
	Simple Q.	**Complex Q.**	**Total**
$F_{m,s,q} \geq \text{median}(\cup_{q,s}\{F_{s,q}\})$	12	9	21
$F_{m,s,q} < \text{median}(\cup_{q,s}\{F_{s,q}\})$	7	10	17
Total	19	19	38

for completeness we list below the respective Fisher's tests results:
Table 43a: $p - value = 0.085$, $CI = [0.91; 4.82]$, $OR = 2.08$.
Table 43b: $p - value = 0.18$, $CI = [0.59319.97]$, $OR = 6.15$.
Table 43c: $p - value = 0.47$, $CI = [0.42; 12.95]$, $OR = 2.18$.
Table 43d: $p - value = 0.51$, $CI = [0.44; 8.47]$, $OR = 1.87$.

8.3.4 Post-task questionnaire

To control the effect of the experiment settings on the results we analyse subjects feedback collected with post-task questionnaire after the application task. Table 44 presents the descriptive statistics of the responses to post-task questionnaire. Responses are on a 5-item Likert scale from 0 (strongly agree) to 4 (strongly disagree). Overall, for both tabular and graphical risk models subjects concluded that the time allocated to complete the task was enough (Q1). Due to the problems with the setup of the graphical and UML comprehension task, the subjects who completed the graphical and UML task were not certain that the comprehension questions are clear (Q4) and they did not experienced difficulty to answer the questions (Q5). For the rest of the settings all subjects reported that they were fine.

Since we provided subjects with electronic version of tabular and graphical risk models, we decided to investigate whether the subjects used search / filtering information in tables and diagrams. Interestingly most of the subjects who used tabular and graphical model (68% and 72% respectively) risk models also used search or filtering information in browser or MS Excel, while only 32% of the subjects who used UML risk model search in PDF.

Table 44: Post Task Questionnaire Results – Experiment 5

Q#	Tabular			Graphical			UML		
	Mean	Med.	sd	Mean	Med.	sd	Mean	Med.	sd
Q1	0.53	0.00	0.96	0.78	0.50	1.00	0.84	1.00	0.96
Q2	0.74	1.00	0.65	1.28	1.00	1.23	1.05	1.00	0.62
Q3	0.42	0.00	0.61	1.22	1.00	1.06	1.00	1.00	1.05
Q4	1.05	1.00	0.91	1.78	2.00	1.17	1.68	1.00	1.20
Q5	0.89	1.00	0.88	1.94	2.00	1.00	1.63	1.00	1.07
Q6	0.74	0.00	1.24	1.44	1.00	1.15	1.26	1.00	1.19
Q7	0.37	0.00	0.76	0.89	1.00	0.76	0.68	1.00	0.58
Q8	0.16	0.00	0.37	0.78	1.00	0.65	0.58	1.00	0.61
Q9	Yes (68%) / No (32%)			Yes (72%) / No (28%)			Yes (32%) / No (68%)		

9 Lessons learnt

The analysis of the Precision and Recall and their combination (F-measure) results found for the Experiments 2 to 5 showed, in general, the major comprehensibility of the tabular model, except for the SID experiment.

Interestingly from the informal discussions that followed the Experiments, the graphical models where always 'a priori' thought as the easiest ones to be understood and to be used to explain a concept, which possibly denotes the higher potentiality of a graphical display of the security model.

As such in this section we try to give a possible explanation of the reasons why, despite this high potential, the graphical model performed worse than the tabular. Again, we started our analysis from the informal discussions with some of the examined people, and organised the findings in the next sections

We found out that the tabular model helps people check whether they gave all the possible answers or not. This is possibly due by the fact that the graphical/tabular representations where given to the subject in a PDF which they had to scroll up and down to find the answers: when you have a table to check you control it up to the end bar, while for the graphical representation the end of your search might be fuzzy.

To search for a numerical evidence of this thesis, we decided to count the incomplete answers, namely the answers where the question has been understood but not all the possible answers have been provided by the tested subject. More specifically we consider an incomplete answer when a person provides only correct answers (i.e. the number of correct items is equal to the total number of answers provided) but do not succeed to provide all the answers expected (i.e. the number of correct items provided is strictly smaller than the number of answers expected). Figure 15 provides the evidence of the results divided by experiment.

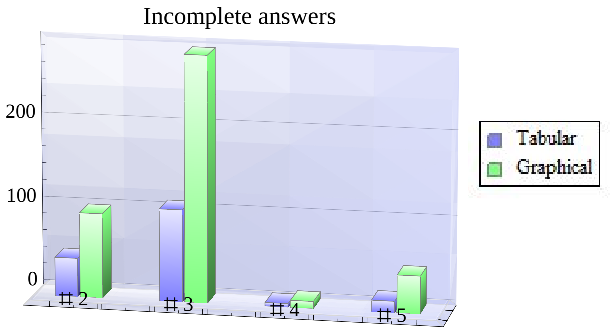


Figure 15: Incomplete answers by experiment

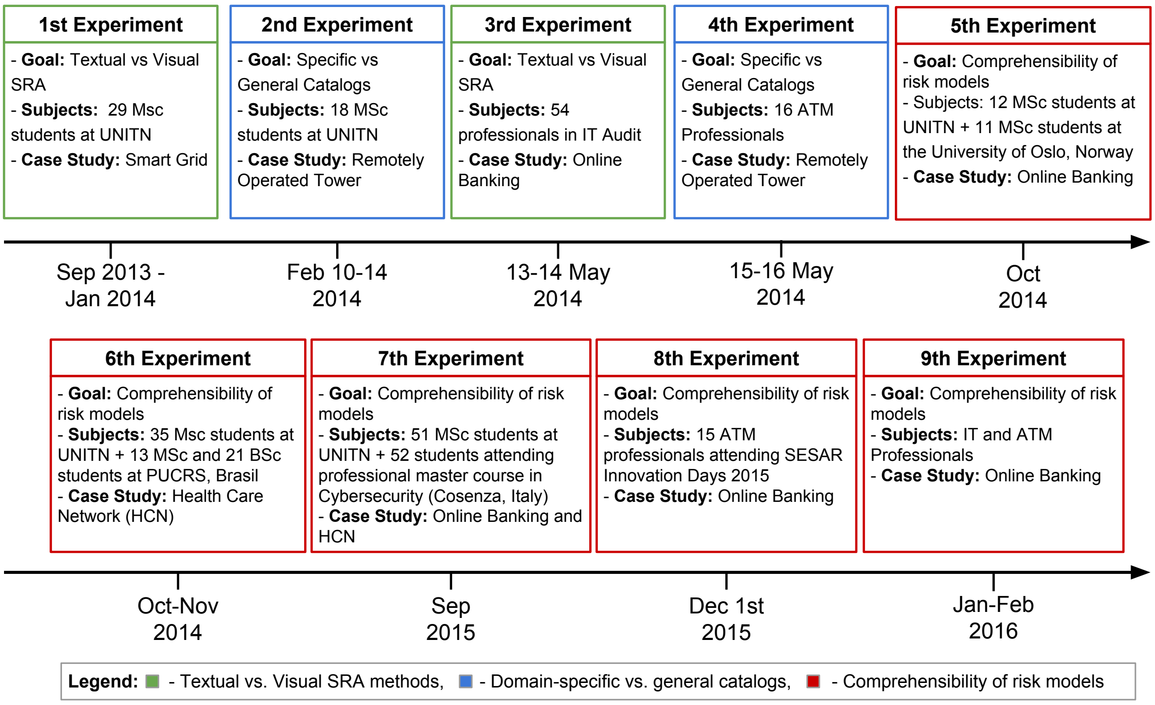
The graphical model counts a remarkably higher number of answers to be considered incomplete by our definition. Which means that the tabular model somehow helps finding in exhaustive way all the solutions.

In order to enhance the graphical risk model and support the identification of all relevant information, we suggest to improve the visualization, search and navigation of graphical risk model through a dedicated tool or query functionality.

10 EMFASE overall result

EMFASE focused its research on three main issues by conducting three types of empirical studies (as shown in Figure 16):

1. The first type aims to evaluate and compare textual and visual methods for security risk assessment with respect to their actual effectiveness in identifying threats and security controls and participants' perception;
2. The second type of studies focuses on assessing the impact of using catalogues of threats and security controls on the actual effectiveness and perception of security risks assessment methods;
3. The third type of studies aims to investigate the comprehensibility of risk models expressed in two modelling approaches: graphical vs. tabular.



1st Experiment
- **Goal:** Textual vs Visual SRA
- **Subjects:** 29 Msc students at UNITN
- **Case Study:** Smart Grid

Sep 2013 - Jan 2014

2nd Experiment
- **Goal:** Specific vs General Catalogs
- **Subjects:** 18 MSc students at UNITN
- **Case Study:** Remotely Operated Tower

Feb 10-14 2014

3rd Experiment
- **Goal:** Textual vs Visual SRA
- **Subjects:** 54 professionals in IT Audit
- **Case Study:** Online Banking

13-14 May 2014

4th Experiment
- **Goal:** Specific vs General Catalogs
- **Subjects:** 16 ATM Professionals
- **Case Study:** Remotely Operated Tower

15-16 May 2014

5th Experiment
- **Goal:** Comprehensibility of risk models
- Subjects: 12 MSc students at UNITN + 11 MSc students at the University of Oslo, Norway
- **Case Study:** Online Banking

Oct 2014

6th Experiment
- **Goal:** Comprehensibility of risk models
- **Subjects:** 35 Msc students at UNITN + 13 MSc and 21 BSc students at PUCRS, Brasil
- **Case Study:** Health Care Network (HCN)

Oct-Nov 2014

7th Experiment
- **Goal:** Comprehensibility of risk models
- **Subjects:** 51 MSc students at UNITN + 52 students attending professional master course in Cybersecurity (Cosenza, Italy)
- **Case Study:** Online Banking and HCN

Sep 2015

8th Experiment
- **Goal:** Comprehensibility of risk models
- **Subjects:** 15 ATM professionals attending SESAR Innovation Days 2015
- **Case Study:** Online Banking

Dec 1st 2015

9th Experiment
- **Goal:** Comprehensibility of risk models
- **Subjects:** IT and ATM Professionals
- **Case Study:** Online Banking

Jan-Feb 2016

Legend: 🟩 - Textual vs. Visual SRA methods, 🟦 - Domain-specific vs. general catalogs, 🟥 - Comprehensibility of risk models

Figure 16: EMFASE experiments

While the first two types of studies have been first conducted with MSc students (Experiment 1 and 2) and then with Professionals (Experiment 3 and 4), the third type has been conducted with MSc students (Experiment 5-7) and with ATM and Security Professionals (Experiments 8 and 9).

The first two types of studies were full scale applications of security risk assessment methods to real-sized application scenario. The subjects had to apply all steps of the method like identification of assets, threats, and security controls. They had to present the results of the risk assessment using tabular or graphical risk modeling notation. While the third type of study the experimentation consisted of experiments, involving both graduate, undergraduate students and ATM and Security Professionals, in which the comprehension task was reading a risk model in either the tabular and graphical representation and answering questions of different complexity about the model. Recall

Table 45: Result overview and way forward

Experiment goal	Experiments	Results	Discussion
To evaluate and compare textual and visual methods for SRA	1 and 3	Textual Methods have higher actual efficacy	- Do not require to learn a new modeling notation - Do not require to learn how to use a tool
		Visual Methods have higher perceived efficacy	- graphical and intuitive representation - A very clear process to identify security risks supported by a dedicated visual tool
To assess the impact of using catalogues of threats and security controls	2 and 4	It was not found "on average" any significant difference in actual efficacy of catalogues	Catalogues can provide a common language for discussion and support in identification of threats and controls but they can limit the envisioning of not already described threats
		Security novices with catalogues performed the same as security experts without catalogues	- Address domain relevant threats also suggesting domain specific controls very useful for novices - Checklists and guidelines with detailed questions for each step of a Security Risk Assessment can be less prescriptive and better support the identification of uncommon and emerging threats and innovative controls by experts
		Domain specific catalogues have higher perceived efficacy	Provide clearer links and a better traceability between threats and controls
To investigate the comprehensibility of graphical or tabular risk models	from 5 to 9	tabular risk models are significantly more effective than the graphical ones with respect to simple comprehension task and slightly more effective for complex comprehension task	- According to Vessey's Cognitive Fit Theory (see D3.2), linear spatial relationships are more easily captured by tabular models - The easiness of searching elements and relationships of tabular risk models is compensated by the easiness of understanding the overall risk picture provided by graphical risk model
		The perceived comprehensibility is the same for both risk modeling notations	The easiness of searching elements and relationships of tabular risk models is compensated by the easiness of understanding the overall risk picture provided by graphical risk model

that according to the MEM (Method Evaluation Model), the Actual Efficacy is the pragmatic success of the method, i.e. the extent to which it improves the performance of the task in question. It can be decomposed in two further terms: the Actual Efficiency, i.e., the effort required to apply a method and the Actual Effectiveness, i.e. the degree to which a method achieves its objectives. The Perceived Efficacy, regards how the method is considered by users, with respect to the following sub-dimensions: the Perceived Ease of Use, i.e., the degree to which a person believes that using a particular method would be free of effort and the Perceived Usefulness, i.e., the degree to which a person believes that a particular method will be effective in achieving its intended objectives (more details on these concepts are reported in D1.2, D1.3 and D3.2 EMFASE deliverables). Table 45 summarises the main findings of all experiments conducted during the project.

Some research questions remain still open, thus further work may be needed to address them. Indeed, we suggest as way forward:

- Real case studies and direct observations of RA methods application by professionals in their work activity and qualitative data gathering;
- Experiments including and evaluating new SRA methods to have a better generalizability of results and new insights.

Bibliography

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

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

[3] EMFASE E.02.32. First empirical evaluation framework. deliverable d1.2, 2014.

[4] EMFASE E.02.32. Selection of risk assessment methods object of study. deliverable d1.1, 2014.

[5] EMFASE E.02.32. Refined empirical evaluation framework. deliverable d1.3, 2015.

[6] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. 39(2):175–191, 2007.

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

[8] Motulsky Harvey. Intuitive biostatistics, 1995.

[9] Natalia Juristo and Ana M Moreno. *Basics of software engineering experimentation*. Springer Publishing Company, Incorporated, 2010.

[10] Barbara A Kitchenham, Shari Lawrence Pfleeger, Lesley M Pickard, Peter W Jones, David C Hoaglin, Khaled El Emam, and Jarrett Rosenberg. Preliminary guidelines for empirical research in software engineering. *Software Engineering, IEEE Transactions on*, 28(8):721–734, 2002.

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

[12] Steve McKillup. *Statistics explained: an introductory guide for life scientists*. Cambridge University Press, 2011.

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

[14] Filippo Ricca, Massimiliano Di Penta, Marco Torchiano, Paolo Tonella, and Mariano Ceccato. The role of experience and ability in comprehension tasks supported by uml stereotypes. volume 7, pages 375–384, 2007.

[15] Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131–164, 2009.

[16] Giuseppe Scanniello, Miroslaw Staron, Hakan Burden, and Rogardt Heldal. On the Effect of Using SysML Requirement Diagrams to Comprehend Requirements: Results from Two Controlled Experiments. pages 433–442, 2014.

[17] Marco Torchiano, Filippo Ricca, and Paolo Tonella. Empirical comparison of graphical and annotation-based re-documentation approaches. *Software, IET*, 4(1):15–31, 2010.

[18] Claes Wohlin, Per Runeson, Martin Hst, Magnus C Ohlsson, Bjrn Regnell, and Anders Wessln. *Experimentation in software engineering*. Springer, 2012.

[19] Robert E Wood. Task complexity: Definition of the construct. 37(1):60–82, 1986.

[20] Robert K Yin. *Case study research: Design and methods*. Sage publications, 2013.