First evaluation report

Document information	
Project Title	EMFASE
Project Number	E.02.32
Project Manager	University of Trento
Deliverable Name	First evaluation report
Deliverable ID	D2.2
Edition	00.01.00
Template Version	03.00.00
Task contributors	
University of Trento; SINTEF; Deep Blue	

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 first 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 way forwards.

Authoring & Approval

Prepared By - *Authors of the document.*		
Name & Company	Position & Title	Date
Martina Ragosta / Deep Blue srl.	Project Contributor	09/12/2014
Alessandra Tedeschi / Deep Blue srl.	WP2 leader	11/12/2014
Federica Paci / UNITN	WP3 Leader	09/01/2015
Katsyarina Labunets /UNITN	Project Contributor	09/01/2015
Martina De Gramatica / UNITN	Project Contributor	14/01/2015

Reviewed By - *Reviewers internal to the project.*		
Name & Company	Position & Title	Date
Bjørnar Solhaug / SINTEF	WP1 Leader	19/01/2015
Elisa Chiarani / UNITN	Project Manager	26/02/2015
Alessandra Tedeschi / Deep Blue srl.	WP2 leader	27/02/2015

Reviewed By - *Other SESAR projects, Airspace Users, staff association, military, Industrial Support, other organisations.*		
Name & Company	Position & Title	Date
<Name / Company>	<Position / Title>	<DD/MM/YYYY>

Approved for submission to the SJU By - *Representatives of the company involved in the project.*		
Name & Company	Position & Title	Date
Fabio Massacci / UNITN	Project Coordinator	<24/03/2015>

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	09/12/2014	First draft	Martina Ragosta	New Document
00.00.02	11/12/2014	Working draft	Alessandra Tedeschi	Internal review
00.00.03	09/01/2015	Working draft	Federica Paci	Review deliverable structure
00.00.04	14/01/2015	Working draft	Martina De Gramatica	Section 8 “Evaluation result overview and discussion” inputs
00.00.05	19/01/2015	Working draft	Bjørnar Solhaug	Review deliverable structure and section 8 “Evaluation result overview and discussion” inputs

00.00.06	28/01/2015	Working draft	Martina Ragosta	Address review comments, overall check and minor changes
00.00.07	30/01/2015	Working draft	Alessandra Tedeschi	Overall check and harmonization before internal release
00.00.08	02/02/2015	Working draft	Martina Ragosta	Document final draft – internal release
00.00.09	26/02/2015	Working draft	Elisa Chiarani	Quality Check
00.00.10	27/02/2015	Final document	Alessandra Tedeschi, Martina Ragosta	Quality check. Finalization of the document for PO feedbacks
00.01.00	24/03/2015	Final document	Fabio Massacci	Approved for official submission

Intellectual Property Rights (foreground)

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

Table of Contents

List of tables

List of figures

Executive summary

The main objective of WP2 of the EMFASE project is to provide support to decision makers for selection of Risk Assessment methods for security in the ATM domain. This support will take the form of guidelines for how to select the risk assessment method best suited for the particular situation (concept under assessment and its maturity level, involved stakeholders, time and budget constraints, etc.).

WP2 empirically evaluates different risk assessment methods in terms of performance, measurable security impact, usability, and economy. The evaluation methods that will be employed in this work package can be case studies and/or controlled experiments, as prescribed by the empirical evaluation framework developed in WP1. During these studies, different risk assessment methods will be applied on different application scenarios.

The design of concrete studies and controlled experiments and their results analysis are relevant objectives of WP2, particularly in its early phases, in order to revise the empirical framework and the set of success criteria, and thereby better adapt the framework to the criteria of significance.

The purpose of D2.2 deliverable is to document the application of the EMFASE empirical evaluation framework to different controlled experiments.

D2.2 further details each experiment from a methodological and procedural point of view, with particular focus on the achieved results and preliminary findings. These provide a first high level sketch of guidelines which can support the redefinition of the EMFASE framework for the empirical evaluation (presented in D1.2 and to be enhanced and detailed in D1.3).

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 [1].

This document presents the first 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 way forwards.

More specifically the document is structured as follows:

- Section 2 presents the experiment framework consisting of the empirical protocol and the experiments overview
- Sections 3 to 7 contain a detailed methodological and procedural description of each controlled experiment conducted so far in. They include also the achieved results, some additional analyses to provide causal explanation of the experiments findings will be presented in D3.1.
- Section 8 offers an overview on the evaluation results and some suggestions for improving and validating the preliminary version of the EMFASE Experimental Framework pre" derived from the experiment result analysis and lessons learnt.

1.2 Intended readership

As stated in Section 1.1, D2.2 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 initial guidelines that are the basis of the next evaluation phase. 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

Term	Definition
AE	Actual Effectiveness
ATM	Air Traffic Management
E-ATMS	European Air Traffic Management System
ITU	Intention To Use
MEM	Method Evaluation Model
MSSC	Minimum Set of Security Controls
OFAs	Operational Focus Areas
OSED	Operational Service and Environment Description
PEOU	Perceived Ease Of Use
PU	Perceived Usefulness
RQs	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 Experiments framework

2.1 Empirical protocol to compare SRA methods

In this section we present a protocol that can be applied to conduct empirical studies to compare two security risk assessment methods with respect to the EMFASE framework scheme and to the success criteria [1]. This protocol was used in conducting the EMFASE experiments 1 to 5. Conceptually, the protocol is divided in two parallel streams that are merged in time as shown in Figure 1.

The **execution stream** is the actual execution of the experiment in which the methods are applied and the experiment results are produced and evaluated. It consists of the following phases: a) **Training**: The participants attend lectures on the industrial application scenarios (**E1**) given by the domain expert, and lectures on the risk assessment method **(E2**) given by the method inventor or by a trusted proxy. **E1** targets the threat to conclusion validity related to the bias that might be introduced by previous knowledge of the participants on the scenario. The domain expert provides to the group a uniform focus and target for the security risk assessment. **E2** limits the threat to internal validity related to the implicit bias that might be introduced by having to train the participants in one's own method as well as a competitor's method; b) **Application:** The participants learn the method by applying it to the application scenario (**E3**) and give a short presentation (**E4**) about the preliminary results**.** These steps address one of the major threats to internal validity, namely that the time spent in training participants is too short for participants to effectively apply the method. The group presentation in **E4** captures a phenomenon present in reality: meeting with customers in order to present progress and gather feedback; c) **Evaluation**: The participants' final reports are collected for evaluating the actual effectiveness of the methods (**E5**).

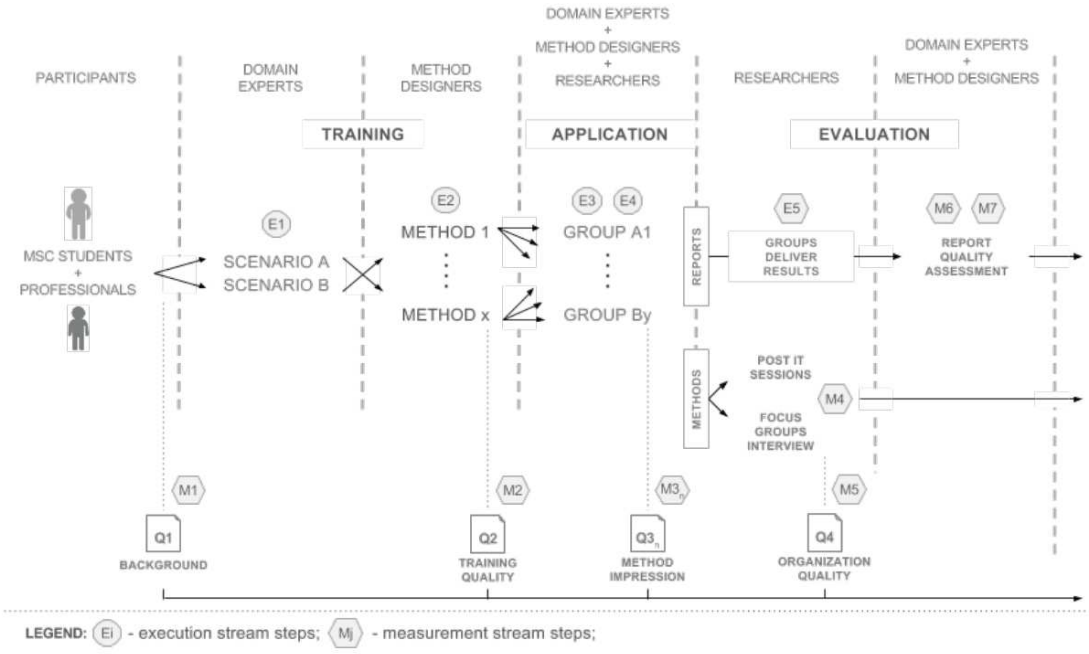

Figure 1: Empirical protocol to compare two SRA methods

The **measurement stream** gathers the quantitative and qualitative data that will be used to evaluate the methods. Similarly to the execution stream, it consists of three phases: a) **Training**: The participants are administered a demographic questionnaire (**M1**). Then**,** participants are distributed a post training questionnaire to determine their initial perception of the methods and the quality of the tutorials (**M2**). M1 targets the threat to internal validity represented by participants' previous knowledge of the other methods; b) **Application**: The participants are requested to answer a post-task questionnaire about their perception of the method after each application session (**M3**); c) **Evaluation**. Participant's perception and feedback on the methods are collected through post-it note sessions, and focus group interviews (**M4**). Participants are also requested to answer a post-task questionnaire about the quality of empirical study's organization (**M5**). Furthermore, the method

designers evaluate whether the groups of participants have applied the method correctly (**M6**), while domain experts assess the quality of identified threats and security controls (**M7**). The last two steps address two issues that may affect both conclusion and construct validity. Indeed, any method can be *effective* if it does not need to deliver useful results for a third party.

2.2 Experiments 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.

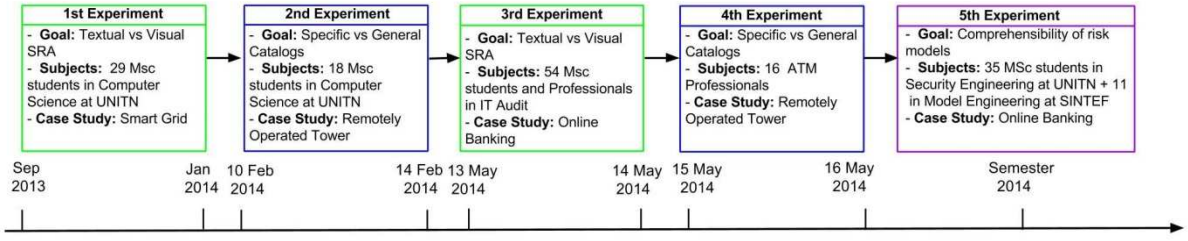

Figure 2: Empirical studies timeline

As shown in Figure 2, we have conducted three types of empirical studies:

1. The first type aims to evaluate and compare textual and visual methods (highlighted in green in Figure 2) 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 (highlighted in blue in Figure 2) 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 (highlighted in violet in Figure 2) expressed in two modelling approaches: graphical vs. tabular.

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), at the moment the third type has been conducted only with MSc students (Experiment 5). Moreover, results of these latest experiments are still under analysis.

Albeit with some variations, i.e. the application case studies, the experiments focus on three Security Risk Assessment methodologies: EUROCONTROL Security Risk Management Toolkit [2], SecRAM [3] and CORAS [4]. We used different Case Studies, drawn from various domains, in order to assess also the applicability and customizability of the Security Risk Assessment Methods under analysis.

EUROCONTROL Security Risk Management Toolkit is an industrial method used to conduct security risk assessment in the air traffic management domain (ATM). It supports the security risk management process for a project initiated by an air navigation service provider, or ATM project, system or facility. EUROCONTROL Security Risk Management Toolkit provides a systematic approach to conduct security risk assessment which consists of five main steps: defining the scope of the system, assessing the impact of a successful attack, estimating the likelihood of a successful attack, assessing the security risk to the organization or project, and defining and agreeing a set of management options. In this method, tables are used to represent the results of each step's execution.

SecRAM is developed within the SESAR JU project 16.02.03 (Security Risk Assessment – Security Risk Assessment Methodology). Its objective is to provide a method that is applicable to all Operational Focus Areas (OFAs), that is understandable to personnel with little expertise and background in security and risk management, and that allows security risk assessment results from different OFAs to be compared. The users are provided with various repositories such as a security register (with lists of assets, threats, threat scenarios, vulnerabilities and controls), security high level documents (including the Minimum Set of Security Controls (MSSC) and security policies), and the Operational Service and Environment Description (OSED). These repositories, along with the

SecRAM Implementation Guidance Material [6], compensates the relative simplicity of the method by providing much of the risk information that otherwise would have to be built from scratch.

CORAS is a model-driven approach to risk assessment that is closely based on the ISO 31000 risk management standard. It consists of three tightly interwoven artefacts, namely the CORAS method, the CORAS language and the CORAS tool. The method follows a process of eight steps that complies with the risk assessment process of the ISO standard. In addition to describing the steps and the activities to be conducted, CORAS comes with practical guidelines and techniques that are needed for carrying out the risk assessment. The language is a graphical notation with various kinds of diagrams that are used throughout the process from beginning to end. While being a formal language with support rigorous analysis of the diagrams, the language was developed to facilitate communication between stakeholders involved in the assessment, including people with little technical background. The CORAS tool is basically a diagram editor for creating all kinds of CORAS diagrams. The tool was designed to facilitate on-the-fly modelling of diagrams during structured brainstorming.

According to the goals of the 1st and the 3rd experiment for evaluating "Textual vs Visual SRA", CORAS was selected as instance of a visual method, and EUROCONTROL Security Risk Management Toolkit and SecRAM as instance of a textual method, respectively.

Moreover, SecRAM comes with catalogs of threats and security controls which have been used as an instance of domain-specific catalogs. For the domain-general catalogues we chose the threats and security controls catalogs of the BSI IT-Grundschutz standard [5]. This standard is developed by Bundesamt für Sicherheit in der Informationstechnik (BSI1), and it is widely used in Germany. It is compatible with the ISO 2700x family of standards. The BSI IT-Grundschutz catalogs not only describe possible threats and what has to be done in general to mitigate them, but they also provide concrete examples on how security controls should be implemented.

These catalogues have been used for evaluating "The Effect of Domain Specific vs Domain General Catalogues" in the 2nd and the 3rd experiment.

Each experiment consists of 3 phases: training, application and evaluation phase. These phases are explained in detail in the following sections according to the specific experiment. However, to each phase corresponds different measurement techniques to collect quantitative and qualitative data, as reported in the following table.

Execution stream	Activities designed	Measurement stream
Training	- Participants attend lectures on the industrial application scenarios by the domain expert; - Participants attend lectures about the method by the method designer	Q1 Background Questionnaire
Application	- Participants work in groups and apply the method to the application scenarios	Q2 Post-Tasks Questionnaire
Evaluation	- Groups deliver a report about the results of the method application and receive feedback	Report evaluated by domain experts

Table 1: Different measurement techniques to collect quantitative and qualitative data according to the experiment phase

As depicted in the previous table, during the training phase the participants are administered a questionnaire with simple factual questions (see Appendix A.1, Q1 Background) to collect information about their background, namely data on their education/work experience and their level of expertise in requirement engineering, security and on other methods they may know. Demographic data are mainly useful to define the participants sample and to face the bias that might be introduced by the previous knowledge of the participants on the methods and on the scenario itself. These data will be analysed through statistical techniques.

At the end of the application phase a post-task questionnaire (see Appendix A.2, Q2 Post-tasks) is distributed to the participants in order to assess feedbacks related to the risk assessment methodology applied. The questionnaire is based on the MEM framework [8] which is a theoretical model that is based on Technology Acceptance Model (TAM) [9], and the Theory of Reasoned Action [10] and the Methodological Pragmatism from the philosophy of science [11].

The resulting theoretical model combines two different but related dimensions of method "success": actual effectiveness and adoption in practice. Actual efficacy is the pragmatic success of the method, i.e. the extent to which it improves the performance of the task in question. Adoption in practice is the extent to which the method is used in practice. These two dimensions are captured by the MEM as summarized in Figure 3. It consists of the following constructs.

- Actual efficiency: The effort required to apply a method;
- Actual effectiveness: The degree to which a method achieves its objectives;
- Perceived ease of use: The degree to which a person believes that using a particular method would be free of effort;
- Perceived usefulness: The degree to which a person believes that a particular method will be effective in achieving its intended objectives;
- Intention to use: The extent to which a person intends to use a particular method;
- Actual usage: The extent to which a method is used in practice.

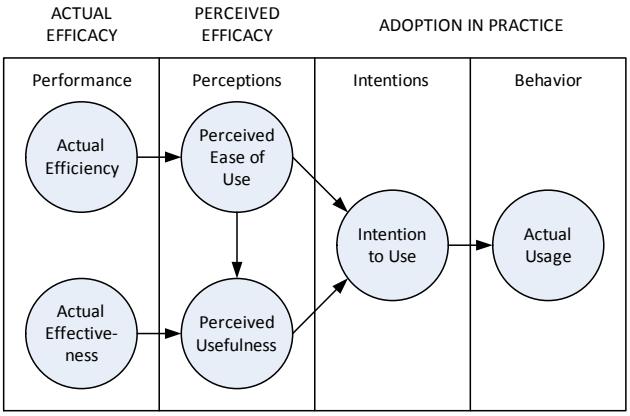

Figure 3: Method Evaluation Model

In MEM, Rescher's theory of Methodological Pragmatism predicts that methods that are more efficient and/or effective in achieving their objectives will be adopted in favour of other methods. This model proposes a slightly different view: those methods will be adopted based on perceptions of their ease of use and usefulness. Actual Efficiency and Effectiveness determine intentions to use a method only via perceptions of ease of use and usefulness. This is a subtle difference, but very important in human behaviour, subjective reality is more important than objective reality. While perceptions of ease of use and usefulness will be partly determined by actual efficacy, they will also be influenced by other factors (e.g. prior knowledge, experience with particular methods, normative influences).

The questionnaire provides data about the perceived ease of use (PEOU), the perceived usefulness (PU) and the intention to use (ITU) measuring the perception through opinion questions that will be analysed on the Likert scale. The actual effectiveness of the methodologies (AE) is assessed based on the number and quality of threats and controls identified by the participants in the Exercise Sheet (see Appendix A.3, Exercise sheet) that participants have to deliver at the end of the experiment. This report will be evaluated by experts (see Appendix A.4, Evaluation Sheet).

Data source	Description		Analysis techniques	Research questions
Q1 Background Questionnaire:	Providing demographical data about the participants (age, education length, work experience), as well as participants' level of expertise in privacy, security and risk assessment methodologies previously applied	Simple factual questions	Statistical	-
Q2 Feedback Questionnaire	Providing an evaluation of the methods' aspects. The questionnaire is based on two different types of questions: - Opinion questions - Open questions	Opinion questions	Likert scale based on MEM study	PEOU, PU and ITU
		Open questions	Coding on pre-defined set of codes	Qualitative explanations
Exercise Sheet	Presenting the results achieved through the risk assessment methodology. For SECRAM an Excel file with threats and controls; for CORAS a PPT file with threats and controls		Counting of threats and controls identified	AE

Table 2: Different analysis techniques to collect quantitative and qualitative data according to the data source

The overall perception on the methods will be assessed through open questions where participants are asked to freely express their opinion. For the analysis of these data coding methodology, drawn from grounded theory [12], will be applied, as well as for knowing the intention to use the methods again. Coding is an interpretive technique that both organizes and supports the interpretation of the data and provides a means to introduce their analysis with quantitative statistical methods. The analytical coding process categorises data to facilitate further qualitative (explanatory) or quantitative (statistical) analyses (also refer to [1] for details). All the raw data collected during the experiments will be analysed through coding techniques

According to the particular experiment, i.e. with the professionals, experiments materials have been adapted with different levels of information details and complexity and data gathering has been improved with interviews and focus groups.

3 1st Experiment: textual vs visual methods for security risk assessment with MSc students

The experiment [7] involved 29 MSc students in Computer Science at the University of Trento (UNITN). They applied EUROCONTROL Security Risk Management Toolkit and CORAS to an application scenario from the Smart Grid domain. The Smart Grid is an electricity network that uses information and communication technologies to optimize the distribution and transmission of electricity from supply points to end-consumers. The application scenario focused on the gathering of metering information from the smart meters located in private households and its communication to the electricity supplier for billing purposes. CORAS was selected as instance of a visual method, and EUROCONTROL Tool kit as instance of a textual method.

3.1 Research method

The goal of the experiment was to compare visual and textual methods for security risk assessment with respect to how successful they are in identifying threats and security controls. For this purpose we have adopted as dependent variables the success constructs defined in the Method Evaluation Model (MEM) proposed by Moody [8]: effectiveness, perceived ease of use, perceived usefulness, and intention to use. Therefore, we have specified the following research questions that match the constructs of the MEM:

- RQ1 Is the effectiveness of the methods significantly different between the two types of methods?
- RQ2 Is the effectiveness of the methods significantly different between the two facets?
- RQ3 Is the participants' overall perception of the method significantly different between the two types of methods?
- RQ4 Is the participants' perceived usefulness of the method significantly different between the two types of methods?
- RQ5 Is the participants' perceived ease of use of the method significantly different between the two types of methods?
- RQ6 Is the participants' intention to use the method significantly different between the two types of methods?

We translated research questions RQ1 - RQ6 into a list of null hypotheses to be statistically tested. We do not list them here due to the lack of space. The interested reader is referred to [7]. To answer RQ1 and RQ2 we measured methods' actual effectiveness by counting the number of threats and security controls identified with each method application and we asked external security experts to assess their quality. Research questions RQ3-RQ6 was investigated by administering to the participants a post-task questionnaire inspired to the MEM after they completed each of the method applications. To gain a better understanding why there is a difference in methods effectiveness and perception we conducted individual interviews with the participants.

3.2 Experimental procedure

We chose a within-subject design where all participants applied both methods to ensure a sufficient number of observations to produce significant conclusions. In order to avoid learning effects, the participants had to identify threats and security controls for different types of security facets of a Smart Grid application scenario. The security facets included Network Security (Network) and Database/Web Application Security (DB/WebApp). For example, for Network Security facet, participants had to identify network security threats like man-in-the-middle attack or DoS attack and proposed security controls to mitigate them.

The participants were randomly assigned to treatments: half of the participants applied first the visual method to network security facet while the second half applied the methods in the opposite order.

The experiment was performed during the Security Engineering course held at University of Trento from September 2013 to January 2014. The experiment was organized in three main phases:

- **Training phase:** The participants were given a 2 hours tutorial on the Smart Grid application scenario (not ATM –related to better test SRA generality and customizability) and a 2 hours tutorial on visual and textual methods. Subsequently the participants were administered a questionnaire to collect information about their background and their previous knowledge of other methods;
- **Application phase**: Once trained on the Smart Grid scenario and the methods, the participants had to repeat the application of the methods on two different facets: Network and Database and Web Application Security. They could deliver intermediate presentations and reports to get further feedback. At the end of the course, each participant submitted a final report documenting the application of the methods on the two facets. For each facet, the participants:
    - Attended a two hours lecture on the threats and possible security controls specific for the facet but not concretely applied to the scenario.
    - Had 2,5 weeks to apply the assigned method to identify threats and security controls specific for the facet.
    - Gave a short presentation about the preliminary results of the method application and received feedback.
    - Had one week to deliver an intermediate report to get feedback.

At the end of the course in mid-January 2014, each participant submitted a final report documenting the application of the methods on the two facets

- **Evaluation phase**: The participants provided feedback on the methods through questionnaires and interviews. After each application phase the participants answered an on-line post-task questionnaire to provide their feedback about the methods. In addition, after final report submission each participant was interviewed for half an hour by one of the experimenters to investigate which are the advantages and disadvantages of the methods

3.3 Results

Since a method is effective based not only on the quantity of results, but also on the quality of the results that it produces, we asked two domain experts to independently evaluate each individual report. To evaluate the quality of threats and security controls the experts used a four item scale: *Unclear* (1), *Generic* (2), *Specific* (3) and *Valuable* (4). We evaluated the actual effectiveness of methods based on the number of threats and security controls that were evaluated as *Specific* or *Valuable* by the experts. In what follows, we will compare the results of all methods' applications with the results of those applications that produce at least specific and valuable threats and security controls.

According to MEM, with regard to the **actual effectiveness**, Figure 4 (top) shows that in this experiment the textual method did better than the visual one in identifying threats. But the results of the Friedman test do not show any significant differences in the number of threats among both all (Friedman test returned *p-value* = 0.57) and specific threats (Skillings– Mack test returned *p-value* = 0.17).

In contrast, Figure 4 (bottom) shows that the visual and textual methods produced the same number of security controls. This is attested also by the results of statistical tests, which show there is no statistically significant difference in the number of security controls of any quality (Friedman test returned p-value = 0.57) and specific security controls (ANOVA test returned p-value =0.72). Thus, we can conclude that in this experiment there was no difference in the actual effectiveness of the visual and textual method for security risk assessment.

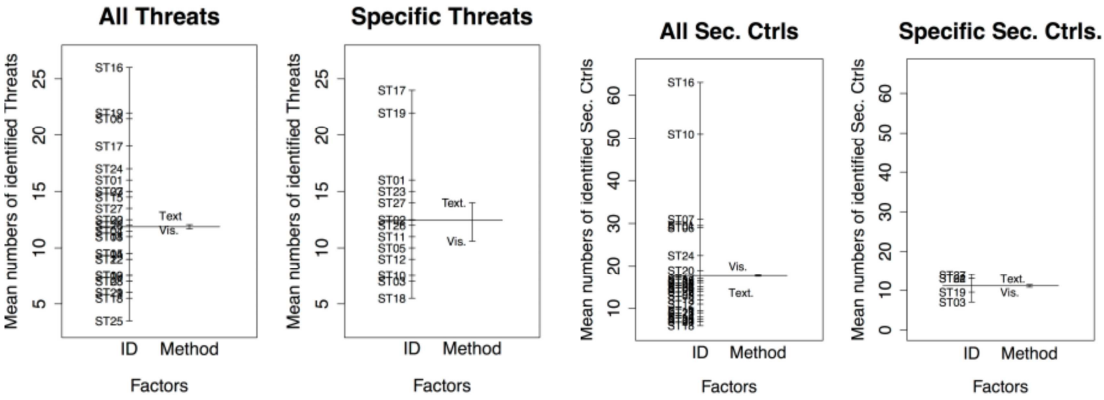

Figure 4: Actual Effectiveness: Number of threats and security controls

With regard to the **participants' perception**, the average of responses shows that the participants preferred the visual method over the textual method with statistical significance (Mann-Whitney test returns Z = -5.24, p-value = 1.4 ∗10−7, es = 0.21).

For what concerns the **perceived ease of use**, the visual method scored better than the textual, and the difference is statistically significant (Mann-Whitney test returns Z= -4.21, p-value = 2 ∗10−5, es = 0.38). But we cannot rely on this result because homogeneity of variance assumption is not met.

Regarding to the **perceived usefulness**, the visual method scored better than the textual with statistical significance (Mann-Whitney test returns Z= -2.39, p-value = 1.7 ∗□10−2, es = 0.15).

For what concerns the **intention to use**, the visual method scored better than the textual with statistical significance (Mann-Whitney test returns Z = -2.05, p-value = 3.9 ∗□10−2, es = 0.16).

Overall we can conclude that in this experiment the visual method is preferred over the textual one with statistical significance.

The difference in the perception of the visual and textual methods can be likely explained by the differences between the two methods. The diagrams of the visual method help participants in identifying threats and security controls because they give an overview of the threats that harm an asset, while using tables makes it difficult to keep the link between assets and threats.

4 2nd Experiment: the effect of domain specific vs domain general catalogues with MSc students

The experiment [13] involved 18 MSc students in Computer Science at UNITN during the EIT ICT Labs Winter School on Secure Design. They were divided into nine groups: half of them applied SESAR SecRAM with the domain-specific catalogues and the other half with the generic catalogues. We chose as instance of domain-general catalogs the threats and security controls catalogs of the BSI IT-Grundschutz standard [5].

Each group had to conduct a security risk assessment of the Remotely Operated Tower (ROT) operational concept. ROT is a technical solution deployed at small and medium-sized airports, which enables an airport tower to be remotely operated via a digital network without human controllers on-site. A set of 360° cameras, sensors and surveillance radars located at the aerodrome provides a 360- degree real -time view of the airports and exhaustive information. This data is used by Air Traffic Control and/or Aerodrome Flight Information Services Operators at Remotely Operated Tower Centres which remotely control different airports simultaneously.

4.1 Research method

The goal of this empirical study from the focus group interviews with ATM professionals, namely the use a catalogue of threats and security controls. Catalogues are documents that contain a list or record of information, such as threats or security controls, arranged in an orderly way and often including descriptions or illustrations. They are widely used by security practitioners as supporting materials for accomplishing security risk assessments, they are also recommended as best practices by national authorities and international agencies. They can enhance the threats and control identifications, but also influence the security expert in its analysis.

In particular we evaluated the effect of using domain-specific and generic catalogues of threats and security controls on the effectiveness and perception of SESAR SecRAM [3]. As for the previous experiment, the comparison was based on the success constructs defined in MEM. Therefore, we specified the following research questions that match the constructs was to evaluate the effect of one of the success criteria that emerged of the MEM:

- RQ1 Is there any difference in the actual effectiveness of the method when used with domain-specific catalogs and with domain-general catalogs?
- RQ2 Is there any difference in participants' overall perception of the method when used with domain-specific catalogs and with domain-general catalogs?
    - RQ2.1 Is there any difference in participants' PEOU of the method when used with domain-specific catalogs and with domain-general catalogs?
    - RQ2.2 Is there any difference in participants' PU of the method when used with domain-specific catalogs and with domain-general catalogs?
    - RQ2.3 Is there any difference in participants' ITU of the method when used with domain-specific catalogs and with domain-general catalogs?
- RQ3 Is there any difference in participants' overall perception of domain-specific catalogs and domain-general catalogs?
    - RQ3.1 Is there any difference in participants' PEOU of domain-specific catalogs and domain-general catalogs?
    - RQ3.2 Is there any difference in participants' PU of domain-specific catalogs and domain-general catalogs?
    - RQ3.3 Is there any difference in participants' ITU of domain-specific catalogs and domain-general catalogs?

We have translated research questions RQ1 – RQ3 into a list of null hypotheses to be statistically tested. To answer RQ1 we measured method's actual effectiveness by counting the number of threats and security controls identified with each method application and by assessing their quality. In fact, if we consider only the number of results but not the quality, threats to conclusion validity may

arise. The participant's perception (RQ2 and RQ3) of the method and catalogs was measured by means of a post-task questionnaire inspired to the MEM. The questions were formulated in opposite statements format with answers on a 5-point Likert scale. To prevent "auto-pilot" answers to our questionnaire half of the questions were given with the most positive response on the left and the most negative on the right while the rest were given in an opposite order. The post-task questionnaire is reported in appendix (see A.2).

4.2 Experimental procedure

We selected SESAR SecRAM as security risk assessment method to be applied by the participants.

The experiment was held in February 2014 and organized in three main phases:

- **Training phase**: The participants were administered a questionnaire to collect information about their background and previous knowledge of other methods. Then they were given a tutorial by a domain expert on the application scenario of the duration of 1 hour. After the tutorial the participants were divided into groups and received one of two sets of catalogues of threats and security controls. The participants were given a tutorial on the method application of the duration of 8 hours spanned over 2 days. The tutorial was divided into different parts. Each part consisted of 45 minutes of training of a couple of steps of the method, followed by 45 minutes of application of the steps and 15 minutes of presentation and discussion of the results with the expert;
- **Application phase**: Once trained on the application scenario and the method, the participants had at most 6 hours in the class to re-perform their security risk assessment with the help of catalogues. After the application phase participants delivered their final reports;
- **Evaluation phase**: Participants were administered a post task questionnaire to collect their perception of the method and the catalogues. Three domain experts assessed the quality of threats and controls identified by the participants

We chose a between-subject design where participants work in groups of two and apply the security risk assessment method with one of two types of catalogues. Nine groups were randomly assigned to treatments: four groups applied security risk assessment method to the ROT scenario using domain-general catalogues while the other five groups used domain-specific catalogues.

4.3 Results

To avoid bias in the evaluation of SESAR SecRAM and of the catalogues, we asked three experts in security of ATM domain to assess the quality of threats and security controls identified by the participants. To evaluate the quality of threats and security controls they used a 5-item scale: Bad (1), when it is not clear which are the final threats or security controls for the scenario; Poor (2), when they are not specific for the scenario; Fair (3), when some of them are related to the scenario; Good (4), when they are related to the scenario; and Excellent (5), when the threats are significant for the scenario or the security controls propose real solutions for the scenario. We evaluated the actual effectiveness of the method used on the catalogues based on the number of threats and security controls that were evaluated Good or Excellent by the experts. In what follows, we will compare the results of all method applications with the results of those applications that produced Good and Excellent threats and security controls.

According to MEM, with regard to the **actual effectiveness**, we analysed the differences in the number of threats identified with each type of catalogue. As shown in Figure 5 (top), there is no difference in the number of all and specific threats identified with each type of catalogues. This result is supported by t-test that returned p-value = 0.8 ($t(7) = 0.26$, Cohen's $d=0.17$) for all threats and p-value = 0.94 ($t(6) = -0.08$, Cohen's $d=0.06$) for specific threats. Figure 5 (bottom) compares the mean of the number of all security controls identified and specific ones. We can see that domain-specific catalogues performed better than domain general catalogues both for all security controls and for specific ones. However, Mann-Whitney test shows that this difference is not statistically significant in case of all security controls ($Z = -0.74$, p-value = 0.56, $r = -0.24$) and specific ones ($Z = -1.15$, p-value = 0.34, $r = -0.41$). We also compared the quality of threats and controls identified with the two types of catalogues. The quality of threats identified with domain-specific catalogue is higher than the one of threats identified with domain-general catalogue. In contrast, the quality of security controls

identified with the support of domain-specific catalogue is lower than the one of controls identified with domain-general catalogue. However, Mann-Whitney test shows that the difference in the quality of identified threats (Z=-0.74, p=0.24, r=0.42) and security controls (Z=0.77, p=0.52, r=0.26) is not statistically significant.

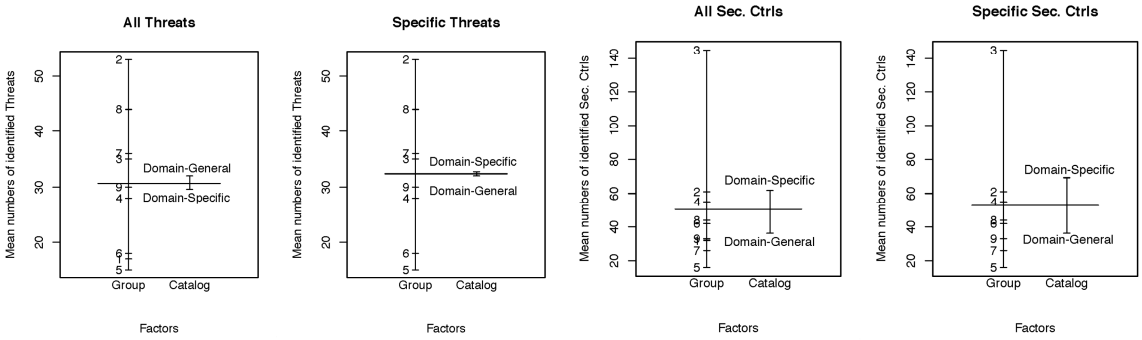

Figure 5: Actual Effectiveness

For what concerns the **method's perception**, the overall perception of the method is higher for the participants that applied domain specific catalogues with statistical significance (Z = -3.97, pvalue = 7 $*$ 10^{-5}, es = 0.17). The same results hold for **Perceived Usefulness** of the method: we have a statistically significant difference (Mann-Whitney test returned: Z = -2.57, p-value = 7.3 $*$ 10^{-3}, es = 0.61) for all participants and good participants (Z = -2.31, p-value = 0.02, es = 0.10).

For **Perceived Ease of Use** and **Intention To Use** the Mann-Whitney test did not reveal any statistically significant difference both for all participants and good participants.

Summarizing, the results indicate that both types of catalogues have no significant effect on the relative effectiveness of the method. In particular, there are no statistically significant differences in the number and quality of threats and security controls identified with the two types of catalogues. However, the overall perception and perceived usefulness of the method is higher when used with the domain-specific catalogues, which are considered easier to use than the domain-general ones.

5 3rd Experiment: textual vs visual methods for security risk assessment with MSc students and professionals

The controlled experiment consisted of a study with 56 students in Computer Science divided in 14 groups, 7 of them applying SecRAM, other 7 CORAS. Every group was supported by the use of the BSI catalogs. A parallel tutorial session was planned in order to train participants on the considered methods and the industrial application scenario (Home Banking) provided by Poste Italiane that focuses on the use of the portal and the online use of prepaid credit card. In particular, Bancoposta is Poste Italiane's banking operations division: it works as a full-fledged bank, providing different services such as bank accounts, credit cards, loans, mortgages and insurance products. In 2001 BancoPosta became the first debit card provider for number of cards issued in Italy and the fifteenth in Europe; the next step in BancoPosta evolution was the creation of the Postepay prepaid debit card which proved to be a huge success for the company. At this moment in time the ever growing number of products offered by BancoPosta coupled with the company's constant interest in exploiting new technologies and developing its business through the online world poses certain issues which cannot be ignored: information security is vital for any financial institution or company dealing with savings deposits, debit or credit cards, customers' accounts must protected from unwanted access and, if this accidentally happens, an immediate response is mandatory.

Each group than had to apply one of the two assigned methods to identify threats and security controls of the real application scenario.

5.1 Research method

The aim of this research was to evaluate the effectiveness of two security risk assessment methods (SESAR SecRAM and CORAS) in identifying threats and security controls towards an application scenario identified in the Poste Italiane's online banking services. This study was conducted through a controlled experiment session involving French MSc students and practitioners attending a Master Course in Audit for Information System Enterprises at Paris Dauphine University.

The evaluation of the methodologies was based on the MEM [8] theoretical framework incorporating constructs to evaluate methods' success referring to their perceived ease of use (PEOU), perceived usefulness (PU), actual effectiveness (AE) and intention to use (ITU).

Four research questions were investigated:

- RQ1: Is the AE significantly different between the methods proposed?
- RQ2: Is the participants' PEOU significantly different between the methods proposed?
- RQ3: Is the participants' PU significantly different between the methods proposed?
- RQ4: Is the participants' ITU the method significantly different between the two types of methods?

From the RQs identified above, corresponding hypotheses to be tested were inferred.

A between-group design was adopted for the experiment, as each group had to apply only one method one the same application scenario. Namely out of 14 groups, 7 of them will apply CORAS and the other 7 SecRAM. Groups were composed by 4 students each. Students were divided according to their studying and working experience, to ensure an equal composition among the groups. The planned division follows:

- 5 groups with 1 M1 student, 2 M2 students and 1 M2 student with working experience;

- 5 groups with 2 M1 students and 2 M2 students;

- 4 groups with 2 M1 student and 2 M2 students with working experience.

5.2 Experimental procedure

The experiment was based on a step-wise process consisting of three interrelating phases: the training session provided by method designers and domain experts first, the application phase in

which participants were requested to apply the methodologies and finally the evaluation phase where the methods' outcomes were assessed.

All these materials was provided to each participant by a USB Key and a paper version.

- **The training phase** was divided in two sub-sessions: firstly a domain expert from Poste Italiane, the Italian provider of postal services, introduced in plenary the main application scenario (BancoPosta Services) that was identified in cooperation with UNITN. This phase was followed by the presentation of SecRAM and CORAS methods.
- The training phase was tightly interconnected with **the application phase**, in a sense that every step introduced through a tutorial given by the method designer was forthwith followed by its application to the case study by the students divided into groups.
- **The evaluation phase** was conducted in several steps: firstly at the end of every application session the groups had to present their intermediate findings, secondly at the conclusion of the experiment activities when students were administered a Questionnaire (Q2). The evaluation of the application was twofold: a post-task questionnaire was administered to participants to collect feedbacks and opinions on the overall effectiveness of the methods (using MEM parameters), while a paper report showed all the results achieved in the risk assessment.

The length of the experiment was two days of continuous work (13th -14th of May 2014).

5.3 Results

The analysis showed that the textual method is slightly more effective than the visual method in the identification of threats, even if this data is not statistically significant. Regarding the security controls, the **Actual Effectiveness** of the textual method is higher than the visual one, since the participants reported a higher number of security controls.

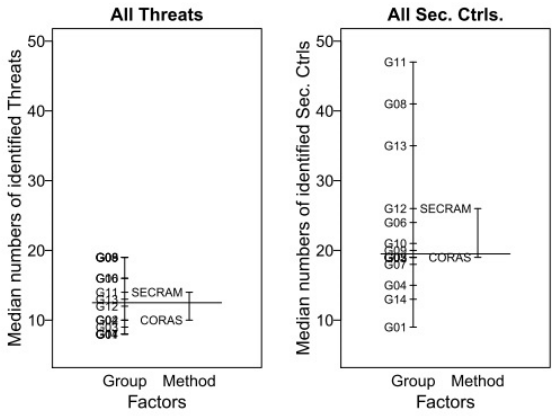

Figure 6: Actual Effectiveness: Number of threats and security controls

The participants revealed a **higher Perceived Ease of Use applying the visual method**, while we found no reliable data for the **Perceived Usefulness** and the **Intention to Use**, due to the failure of the homogeneity variance assumption.

Although the textual method was more effective in the identification of threats and security controls, participants expressed their **Overall Preference towards the visual method** over the textual method.

Summarizing, according to the participants involved in the experiment the textual method performed better in the identification of the security controls, while the visual and the textual method produced similar results in the threats identification.

6 4th Experiment: the effect of domain specific vs domain general catalogues with professionals

The experiment [14] involved 26 professionals from several ATM Italian companies (11 participants were from DeepBlue S.r.l. and 10 from IDS, SESM, SICTA, ASTER and ENAV). The participants were randomly divided in two groups composed by 9 professionals in each (group A and B), and one group composed by 8 professionals (group C). Each group worked individually. The participants belonging to Group A applied the method with the support of EUROCONTROL catalogs; the participants in Group B used BSI catalogs, and the participants in the Group C did not receive any support.

Each group had to apply the method on the same scenario, namely the Remote Operate Tower (ROT) provided by EUROCONTROL (presented in Section 4 and also in [15]). The participants had access to all the materials needed, by an individual reading of the ROT description and by a SecRAM tutorial provided by the method designer (Trainer at the EUROCONTROL IANS).

6.1 Research method

The aim of this controlled experiment organized within the EMFASE project was two-fold:

a) evaluate the effectiveness of SESAR SecRAM risk assessment method in identifying threats and security controls;

b) assess effect that the use of a catalog of threats and security controls has on SESAR SecRAM effectiveness. In particular, we will compare the effect that domain specific catalogs (EUROCONTROL) vs general domain catalogs (BSI).

The evaluation of the methodology and effect of the use of the catalogues was based on the Method Evaluation Model [8], a theoretical framework incorporating constructs to evaluate methods' success referring to their perceived ease of use (PEOU), perceived usefulness (PU), actual effectiveness (AE) and intention to use (ITU). Therefore, we have specified the following research questions that match the constructs of the MEM:

- RQ1: Is there any difference in the number of threats and security controls identified with domain-specific catalogs and with domain-general catalogs (AE)?
- RQ2: Is there any difference in participants' overall perception of using security risk assessment method with domain specific catalogs and with domain-general catalogs?
    - RQ2a: Is there any difference in participants' PEOU of using security risk assessment method with domain-specific catalogs and with domain-general catalogs?
    - RQ2b: Is there any difference in participants' PU of using security risk assessment method with domain-specific catalogs and with domain-general catalogs?
    - RQ2c Is there any difference in participants' ITU of using security risk assessment method with domain-specific catalogs and with domain-general catalogs?
- RQ3: Is there any difference in participants' overall perception between using domain-specific catalogs and domain general catalogs?
    - RQ3a: Is there any difference in participants' PEOU between using domain-specific catalogs and domain-general catalogs?
    - RQ3b: Is there any difference in participants' PU between using domain-specific catalogs and domain-general catalogs?
    - RQ3c: Is there any difference in participants' ITU between using domain-specific catalogs and domain-general catalogs?

From the research questions identified above, corresponding hypotheses to be tested were inferred. The evaluation of the application was twofold: a Post-Tasks Questionnaire, a Post-It Notes session and a Focus Group discussion were used to collect feedback and opinions on the overall effectiveness of the methods (using MEM parameters).

6.2 Experimental procedure

A between-subjects design was adopted for the experiment, as participants work individually and have to apply the method under different conditions. Namely out of 26 participants:

- Group A: 9 participants applied SecRAM with the support of EUROCONTROL catalogs,
- Group B: 9 participants applied SecRAM with the support of BSI catalogs;
- Group C: 8 participants did not use any catalog.

The experiment was held in Rome, from 15th to 16th of May 2014 and was based on a step-wise process consisting of three interrelating phases: training, application and evaluation.

- **The training phase:** The participants were administered a questionnaire to collect information about their background and previous knowledge of other methods. The training was provided by a self-training phase in which participants were asked to read the application scenario description and by a frontal-training phase in which the method designer Dr.Rainer Koeller from EUROCONTROL briefly introduced the SecRAM methodology process through a tutorial. Each method step introduced is forthwith applied on the case study and finally the results achieved in the risk assessment were evaluated during the last phase.
- **The application phase:** Once trained on the application scenario and the method, the participants had time in the class to perform their security risk assessment with or without the help of catalogues. After the application phase participants delivered their security risk assessment
- **The evaluation phase:** Professionals were administered a post task questionnaire in order to assess feedback related to the effectiveness of SecRAM application and the effectiveness of the catalogs used. In addition, the overall perception on the method was assessed through several open questions where participants were asked to freely express their opinion. After the administration of the questionnaire, a Post-It Notes session was conducted in order to collect shared feedbacks and comments from the participants on the principal aspects of the method. Each participant had to fill in 10 Post-It: 5 Post-it with positive aspects of the method and 5 Post-It with negative aspects. Then they were divided into groups constituted according to the catalogs and discussed how to cluster the Post-It in aspect categories and prioritize them from the most relevant to the less significant. Moreover, a focus group session was planned in which participants was asked some questions with the aim to collect qualitative explanations about the categorization previously set. This phase was audio recorded and the audio transcripts will be analyzed through the coding methodology. The actual effectiveness of the methodologies (AE) was evaluated by the domain expert for asserting the whole application of the method toward the ROT scenario, counting the threats and security controls identified by the participants and evaluating them in term of quality. The results achieved through the application phase were collected in a template results that participants had to deliver at the end of the experiment.

6.3 Results

The gathered data were analyzed qualitatively and quantitatively. According to MEM, with regard to the **actual effectiveness**, we measured method's actual effectiveness as a quality of threats and security controls identified by the participants. Two ATM security experts independently assessed the quality. They reported a similar assessment for each group. Figure 7 illustrates the average of experts' evaluation for threats (reported on x-axis) and security controls (on y-axis). Six participants out of fifteen performed poorly. In terms of the final assessment we observed that: a) the experts marked bad participants the same way; b) they consistently marked moderately good participants; and c) they had a different evaluation only for the threats of one participant and for the security controls of another participant out of 15 participants.

We used Wilcoxon test to validate if the difference in experts' evaluation is statistically significant. The results showed that there is no statistically significant differences in the evaluations of two experts both for threats (p = 0:09) and controls (p = 0:77). Therefore, we can conclude that there is no significant effect of treatments on method's actual effectiveness.

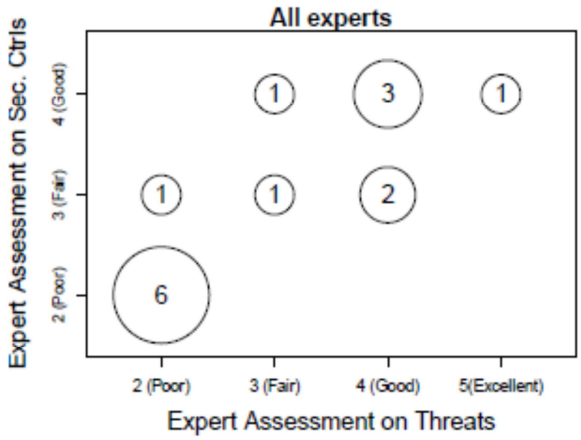

Figure 7: Experts assessment of quality of threats and security controls

Also regarding to the **perceived effectiveness**, there is no difference in method's PU when the method applied with or without catalogues of threats. Same results we have for method's PU regarding security controls identification. Considering method's PEOU, the participants conducted threats identification with domain-general catalogue of threats or without catalogue reported higher method's PEOU than participants applied domain-specific catalogue. While for method's PEOU for security controls identification only the participants conducted risk assessment without catalogues reported higher perception. Therefore, we can conclude that in this experiment there was no significant effect of treatments on method's perceived effectiveness.

With regard to the qualitative results, we can see that both catalogues were easy to use and could be easily applied to a different context, but domain-general catalogues had better perception with respect to questions "finding specific threats/security controls for a different context would be easy with catalogue of threats/security controls". In contrast, domain-specific catalogues had better PU. Participants were not confident about catalogues' usefulness related to "finding specific security controls is more quickly with catalogue of security controls". But they were confident that domain-specific catalogues accelerates finding specific threats and makes participants more productive in the identification of threats and security controls. We clarified these summary findings from the results of focus groups interviews with the participants and post-it notes sessions summarizing each discussion within groups.

7 5th Experiment: comprehensibility of risk models

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 experiments with MSc students. The first experiment 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

7.1 Research method

To study the comprehensibility of risk models expressed in two modeling approaches we wanted to investigate the following research questions:

- RQ1: Which risk model, the graphical one or the tabular one, is easier to understand for participants?
- RQ2: Which risk model, the graphical one or the tabular one, requires less effort from the participants to achieve comprehension?
- RQ3: Which risk model, the graphical one or the tabular one, results in higher productivity of the participants (derived from the ratio between comprehension level and effort?
- RQ4: Which risk model, the graphical one or the tabular one, results in higher perceived comprehension by participants?

The above RQs are addressed by testing the following hypotheses in our experiment:

RQ	H_0	H_A
RQ1: Comprehension	H_{01}: $Compr_{tabular} = Compr_{graph}$	H_{A1}: $Compr_{tabular} \neq Compr_{graph}$
RQ2: Effort	H_{02}: $Effort_{tabular} = Effort_{graph}$	H_{A2}: $Effort_{tabular} \neq Effort_{graph}$
RQ3: Productivity	H_{03}: $Product_{tabular} = Product_{graph}$	H_{A3}: $Product_{tabular} \neq Product_{graph}$
RQ4: Perceived comprehension	H_{04}: $PCompr_{tabular} = PCompr_{graph}$	H_{A4}: $PCompr_{tabular} \neq PCompr_{graph}$

Table 3: Hypotheses to be tested in the "Comprehensibility of risk models" experiment

According to the RQs, the comprehension level (Compr) can be measured as precision and recall for each answer in the comprehension questionnaire. Similar to De Lucia et al. [16], the open questions allowed us to evaluate the answers using Information Retrieval metrics, namely precision and recall:

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

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

where $answer_{s,i}$ is the set of answers given by participants to question i; $correct_i$ is the set of correct answers for question i. To evaluate the average comprehension level we decided to use a measure that aggregates both precision and recall, i.e. F-measure [17]:

$$F - measure_s = 2 * \frac{precision_s * recall_s}{precision_s + recall_s}\%,$$

where

$$recall_s = \frac{\sum_i |answer_{s,i} \cap correct_i|}{\sum_i |correct_i|}\%,$$

$$precision_s = \frac{\sum_i |answer_{s,i} \cap correct_i|}{\sum_i |answer_{s,i}|} \%.$$

The effort can be measured as the time required by a participant to answer the questions in the comprehension questionnaire.

The Productivity (Product) can be measured as $F - measure_s / Effort\ \%$.

Perceived comprehensibility (PCompr) can be measured as the participants' opinion regarding the ease of understanding the models on a 5-point Likert scale.

7.2 Experimental procedure

The population of this control experiment was 35 MSc students enrolled to the Security Engineering course during fall 2014 semester at the University of Trento and 11 MSc students of the Model Engineering course held at the University of Oslo.

We selected as an instance of tabular risk model the tabular representation used by NIST 800-30 standard. This standard has been proposed by National Institute of Standards and Technology and it is open source. The risk assessment process used in NIST standard consists of nine steps. The results of each step execution are documented by mean of tables. As an instance of visual risk model we choose the diagrams used in CORAS method.

We selected an application scenario from the Online Banking domain developed by Poste Italiane. It focused on online banking services provided by Poste Italiane's division through Home Banking Portal, Mobile Application and Prepaid Cards. Based on this scenario we developed two risk models that reported at the end of this document.

The experiment was based on a step-wise process consisting of three interrelating phases: training, application and evaluation. At the beginning of the experiment all participants answered demographics and background questionnaire.

- **Training phase**: All participants attended short 10 minutes tutorial about both types of risk models and application scenario. After, the participants were randomly assigned to one of two tasks orders, so a half of the participants did task related to graphical risk model, the other half did the task related to tabular risk model.
- **Application phase**: During the Application phase the participants were asked to review proposed graphical or tabular risk models and answer the online comprehension questionnaire. Based on the results of the pilot study we found out that 20 minutes is enough time to do the task. Therefore, we limit time of the Application phase with 20 minutes. To avoid learning effects, we assigned participants to the treatments in even/odd order, i.e. each two participants that sat next to each other were assigned to different treatments.
- **Evaluation phase**: After completion of the task, the participants were redirected to a post-task questionnaire about adequacy of the tasks and their perception whether the risk model is easy to understand.

7.3 Results

With regard to this experiment, results are not available yet because the analysis of the gathered data is still in progress. More details about the results of this first Comprehensibility Experiment and of other experiments about Comprehensibility, already carried out in 2014 and planned for 2015, will be provided in D2.3.

8 Evaluation result overview, discussion and way forwards

The results of the empirical studies reported in this deliverable will serve as a basis for further developing the EMFASE empirical evaluation framework [1]. The evaluation framework is currently in its initial version, based on a set of success criteria for SRA methods in the ATM domain. The results of the EMAFSE empirical studies give insight into which criteria actually do have a significant effect on the success of a method. As part of revising the empirical framework we will use the experiment results to revise the set of success criteria, and thereby better adapt the framework to the criteria of significance.

To summarise the main findings from the experiments, we can draw some preliminary conclusions that will inform further project activities in WP1 and WP2.

Regarding the difference between “textual vs visual” methods, analysed in Experiments 1 and 3, it has been showed that Textual Methods have higher actual efficacy, since they do not require to learn a new modeling notation, that may be difficult, and moreover they do not require to learn how to use a tool, that may be very time and effort consuming.

On the other hand, Visual Methods have higher perceived efficacy due to their graphical representation and the Visual Method under analysis (i.e, CORAS) has a very clear process to identify security risks supported by a dedicated visual tool.

Regarding the difference between “domain specific vs domain generic catalogs**”,** analysed in Experiments 2 and 4, main findings were that it was not found ‘on average’ any significant difference in actual efficacy of catalogs. While security novices with catalogs performed the same as security experts without catalogs. This is really interesting and should be better interpreted and further discussed with domain experts.

Domain specific catalogs have higher perceived efficacy, since they are easier to navigate, are written in the ‘domain specific language’ and address domain relevant threats also suggesting domain specific controls. Domain specific catalogs provide clearer links and a better traceability between threats and controls.

In general catalogs can provide a common language for discussion among security experts involved in the Risk Assessment and they can be used to check completeness of results. It was asked by professionals participating to the experiments if also a detailed checklist or appropriate guidelines with relevant ‘questions’ about threats and control can be used in a similar (but perhaps even more effective) way as catalogs. 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.

Finally our experiments have some limitations and “threats to their validity” that should be solved in the future EMFASE round of experiments.

Regarding the Internal validity there is the bias of the different background and expertise of participants. Previous knowledge of participants cannot be eliminated and difficultly eliminated. There are also some problems with respect to the validity of our results, due to the poor statistical significance of the current version of experiments and to the difficult generalization of our results.

Thus, there are still some remaining open issues such as “How long should be an empirical study?”, “How to collect data?”, and “How to overcome language gaps?”. The presented controlled experiments and the derived lessons learnt have provided some potential solutions (e.g. make the experiments at least two days long, try to have much more participants also by having short and remote experiments, provide support to the participants in terms of tools to simplify self-reporting and a mediator to overcome language gaps) that will be adopted in the next evaluation phase.

The results and the experiences from the empirical studies can support the redefinition of the EMFASE framework for the empirical evaluation (presented in D1.3). The EMFASE empirical evaluation framework includes guidelines, not only on what to investigate, but also on how to conduct the empirical studies. Because we conduct the experiments in accordance with the framework scheme and protocol (see [1]), we make use of the lessons learned in better understanding how the scheme and protocol should be designed.

The first Empirical Framework should be improved and validated:

- with the continuous support of Subject Matter Experts and a further review of identified criteria and of their mapping with MEM constructs [1] [8],
- through additional experiments, such as:
    - Real case studies and direct observations of RA methods application by professionals in their work activity and qualitative data gathering (hopefully in collaboration with SESAR P16.06.02 or through SINTEF Security Risk Assessmnet Experts);
    - Experiments organised at IANS and IATA courses in order to collect more feedback from security professionals in the aviation domain;
    - Brief on-line experiments to be distributed among security professionals from various domains, to reach higher number of participants and have statistical significance;
    - Experiments including and evaluating new SRA to have a more generalizability of results and new insights.
    - Experiments for new 'research questions' that will analyse other aspects of the Security Risk Assessment Methods.

After the validation process, we will obtain the final version of the EMFASE Empirical Evaluation Framework including guidelines for ATM security stakeholders. These can be adopted for choosing the right SRA based on factors such as the type of system, the skills and training of the analysts and developers, the roles of involved stakeholders, and the economy involved. The guidelines will be based on scientific methods for empirical evaluation and theories for risk assessment. Empirical guidelines like these for the selection of risk assessment methods have never been developed before.

The results and the experiences from the empirical studies have the potential also to better understand how to build a security case in the development process of an ATM system or solution. Such a security case is envisaged, but still missing, for the E-OCVM [18]. The SESAR JU develops guidance and support for building a security case [19]. The results of the EMFASE experiments and the empirical framework should aid SESAR stakeholders in selecting the SRA methods that are best suitable for building the security case.

9 References

[1] EMFASE, First Empirical Evaluation Framework and Existing Practices, Deliverable D1.2, Edition 00.01.02, 2014

[2] EUROCONTROL, “ATM security risk management toolkit – Guidance material”, 2010

[3] SESAR 16.02.03: SESAR ATM security risk assessment method, Deliverable D02, 2013

[4] Lund, M. S., Solhaug, B., Stølen, K. (2011) “Model-Driven Risk Analysis – The CORAS Approach”, Springer, 2011

[5] BSI 2005. IT-Grundschutz Catalogues.

[6] SESAR Joint Undertaking: SESAR ATM SecRAM implementation guidance material. Project deliverable 16.02.03-D03 (2013)

[7] Labunets, K.; Paci, F.; Massacci, F.; Ruprai, R. (2014) "An experiment on comparing textual vs. visual industrial methods for security risk assessment", Fourth International Workshop on Empirical Requirements Engineering (EmpiRE), pp.28-35

[8] Moody, D. L. (2003) “The method evaluation model: a theoretical model for validating information systems design methods”, In Proc. of ECIS ’03, pp. 1327–1336.

[9] Davis, F. D. (1989), "Perceived usefulness, perceived ease of use, and user acceptance of information technology", MIS Quarterly 13 (3): 319–340.

[10]Madden, Thomas J., Pamela Scholder Ellen, and Icek Ajzen. "A comparison of the theory of planned behavior and the theory of reasoned action." Personality and social psychology Bulletin 18.1 (1992): 3-9.

[11]Rescher, Nicholas. "Methodological pragmatism: A systems-theoretic approach to the theory of knowledge." (1977).

[12]A. L. Strauss and Juliet M. Corbin: Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. SAGE Publications (1998)

[13]Labunets, K., Paci, F., Massacci, F. (2015) “Evaluating the Effect of Using Catalogues on Identification of Security Threats and Controls”, International Symposium on Engineering Secure Software and Systems 2015 (submitted)

[14]de Gramatica, M., Labunets, K., Massacci, F., Paci, F., Tedeschi, A. (2015) “The Role of Catalogues of Threats and Security Controls in Security Risk Assessment: An Empirical Study with ATM Professionals”

[15]EMFASE, Scenario descriptions and requests for EUROCONTROL input, Deliverable D2.1, Edition 00.00.07, 2014

[16]De Lucia, A., Gravino, C., Oliveto, R., Tortora, G. (2010) “An experimental comparison of ER and UML class diagrams for data modelling”, Empirical Software Engineering, 15(5), 455-492.

[17]Baeza-Yates, R., A., Ribeiro-Neto, B. (1999). “Modern Information Retrieval”, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[18]EUROCONTROL: European Operational Concept Validation Methodology (E-OCVM) 3.0, Volume I, 2010

[19]SESAR 16.06.02: SESAR ATM Security Reference Material – Level 1, Deliverable D101, 2013

Appendix A Questionnaires

A.1 Q1 Background

Q1 - Participants Background and Security Awareness

This questionnaire is to collect data about the background of the participants. The answers to this questionnaire are NOT used by any means to evaluate/grade them.

First name: __

Last name: __

1. **How old are you?** _____
2. **What is your gender?** □ Male □ Female
3. **What is your occupation?**

You are currently:

□ only studying, full time
□ also working as an employee
□ also working as a self employed (e.g. consultant)
□ also running your own company
□ other: _____

4. **What is length of your education ?**

Please specify the length of your education in years after your high school degree (i.e. a number of years of university education or professional trainings)

_____

5. **What are your areas of study?**

Please specify your areas of study

_____

6. **Do you have any working experience?**

Please specify the length of your working experience in years

_____

7. **What are you roles at work?**

Answer this question only if at Question 6 you have specified a length of working experience

_____

8. **Have you ever been involved in any Security and Privacy Initiative/Project?**

Please choose only one of the following options □ Yes □ No

9. **Which was your role in the Initiative/Project?**

Please specify your role in the Initiative/Project

_____

_____

A.2 Q2 Post-tasks

Q2 - Method Assessment

This questionnaire is to collect your impressions about the method that you have applied in the first assignment. The answers to this questionnaire are NOT used by any means to evaluate/grade them.

First name: __

Last name: __

Part I - Method (31 questions)

Read questions carefully. The positive and negative statements of the questions are mixed.
The questionnaire has an opposing statements format, so
If you agree strongly with the statement on the left, **check** the leftmost box **(1)**.
If you agree, but less strongly, with the **left statement**, **check** box #2 from the left **(2)**.
If you agree with neither statement, or find them equally correct, **check** the middle box **(3)**.
If you agree, but less strongly, with the **right statement**, **check** box #2 from the right **(4)**.
If you agree strongly with the **statement** on the **right**, **check** the rightmost box **(5)**.

N		1 2 3 4 5	
1.	The method defines the **right level of granularity** of asset, security risk and security control.	☐ ☐ ☐ ☐ ☐	The method defines the **wrong level of granularity** of asset, security risk and security control.
2.	A **catalog of threats** would have made **harder** the identification of threats with this method.	☐ ☐ ☐ ☐ ☐	A **catalog of threats** would have made **easier** the identification of threats with this method.
3.	A **catalog of security controls** would have made **easier** the identification of security controls with this method.	☐ ☐ ☐ ☐ ☐	A **catalog of security controls** would have made **harder** the identification of security controls with this method.
4.	Overall, I think the method **provide** an effective solution to the identification of **security risks**	☐ ☐ ☐ ☐ ☐	Overall, I think the method **does not provide** an effective solution to the identification of **security risks**
5.	Overall, I think the method **does not provide** an effective solution to the identification of **security controls**	☐ ☐ ☐ ☐ ☐	Overall, I think the method **provides** an effective solution to the identification of **security controls**
6.	I found the method **easy to adopt and use** to different contexts.	☐ ☐ ☐ ☐ ☐	I found the method **hard to adopt and use** to different contexts.
7.	Overall, I found this method **difficult to use**	☐ ☐ ☐ ☐ ☐	Overall, I found this method **easy to use**
8.	Overall, I found this method to be **useless**	☐ ☐ ☐ ☐ ☐	Overall, I found this method to be **useful**
N		1 2 3 4 5	

A.3 Exercise sheet

Table Summary of Risk Assessment				
Asset	Incident	Likelihood	Consequence	Risk Level

Overall Summary			
Asset	Risk	Risk Level	Treatments

A.4 Evaluation Sheet

Report Quality Assessment by Domain Experts				
Link to folder with reports:				
Listed below are the criteria and marks followed by domain experts for the report quality assessment:				
	Scale	Threats Quality	Security Controls Quality	
	1 - Bad	Not clear which are the final threats for the scenario	Not clear which are the final security controls for the scenario	
	2 - Poor	Threats are present but are not specific for the scenario	Security controls are present but are not specific for the scenario	
	3 - Fair	Threats are present and SOME of them are related to the scenario	Security controls are present and SOME of them are related to the scenario	
	4 - Good	Threats are present and they are related to the scenario	Security controls are present and they are related to the scenario	
	5 - Excellent	Threats are present and they are major threats for the scenario	Security controls are present and propose real solutions for the scenario	
On the page "Assessment Form" you will find the table with the following columns:				
Group ID		Identifier of group		
Please provide your assessment of each report in the following fields according to the scales presented above:				
Threats Quality		Threats quality inline with the corresponding scale		
Security Controls Quality		Security Controls quality inline with the corresponding scale		
Comments		Here you can provide participants and us your comments to the report		

-END OF DOCUMENT-