Refined Empirical Evaluation Framework

Document information	
Project Title	Empirical Framework for Security Design and Economic Trade-Off – EMFASE
Project Number	E.02.32
Project Manager	University of Trento
Deliverable Name	Refined Empirical Evaluation Framework
Deliverable ID	D1.3
Edition	00.00.01
Template Version	03.00.00
Task contributors	
SINTEF; University of Trento; Deep Blue	

Abstract

This deliverable reports on the refined EMFASE framework for evaluation of methods for security risk assessment in the ATM domain. The presented results are the advances made based on the recent experiments conducted by the project. The current empirical framework is based on the initial version that was documented in deliverable D1.2. It is based on a method evaluation framework (MEM) for evaluating the success of a method, and includes a scheme for conducting empirical studies that incorporates identified success criteria for security risk assessment in the ATM domain. The deliverable also includes an overview and evaluation summary of the EMFASE experiments.

Authoring & Approval

Prepared By - *Authors of the document.*		
Name & Company	Position & Title	Date
Ketil Stølen	WP1 leader	01/08/2015
Alessandra Tedeschi (DBL)	Project member	07/08/2015
Gencer Erdogan (SINTEF)	Project member	14/08/2015
Bjørnar Solhaug	Project member	28/08/2015
Katsiaryna Labunets (UNITN)	Project member	02/09/2015

Reviewed By - *Reviewers internal to the project.*		
Name & Company	Position & Title	Date
<Name / Company>	<Position / Title>	<DD/MM/YYYY>

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
<Name / Company>	<Position / Title>	<DD/MM/YYYY>

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	01/08/2015	Working document	K. Stølen, G. Erdogan	First full draft
00.00.02	07/08/2015	Working document	A. Tedeschi	Section 5
00.00.03	14/08/2015	Working document	G. Erdogan	Sections 2-4
00.00.04	28/08/2015	Working document	B. Solhaug	Revision of all sections
00.00.05	02/09/2015	Working document	K. Labunets	Revision of section 5.3.3

Intellectual Property Rights (foreground)

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

Table of Contents

EXECUTIVE SUMMARY **5**

1 INTRODUCTION **6**

1.1 PURPOSE OF THE DOCUMENT 6
1.2 INTENDED READERSHIP 6
1.3 INPUTS FROM OTHER PROJECTS 6
1.4 GLOSSARY OF TERMS 7
1.5 ACRONYMS AND TERMINOLOGY 7

2 BEST PRACTICES ON EMPIRICAL METHODS **8**

3 SUCCESS CRITERIA FOR ATM SECURITY RISK ASSESSMENT METHODS **9**

3.1 IDENTIFIED SUCCESS CRITERIA 9
3.2 SUCCESS CRITERIA AND RISK ASSESSMENT MEM 11

4 THE EMFASE FRAMEWORK **14**

4.1 PURPOSE AND TARGET GROUP 14
4.2 EMPIRICAL FRAMEWORK 14
4.2.1 Framework Scheme *14*
4.2.2 An Empirical Protocol to Compare Two SRA Methods *16*

5 OVERVIEW OF EMFASE EMPIRICAL STUDIES **20**

5.1 EVALUATING AND COMPARING VISUAL AND TEXTUAL METHODS 21
5.1.1 Experimental Procedure *21*
5.1.2 Experimental Results *22*
5.2 EVALUATING THE EFFECT OF USING CATALOGUES OF THREATS AND CONTROLS 23
5.2.1 Experimental Procedure *23*
5.2.2 Experimental Results *24*
5.3 EVALUATION AND LESSONS LEARNED 26
5.3.1 Success Criteria *26*
5.3.2 Mapping MEM Constructs to Success Criteria *27*
5.3.3 Review the Experimental Protocol *27*

6 CONCLUSION **29**

7 REFERENCES **30**

List of tables

Table 1: Case study research process 8
Table 2: Occurrences of reported success criteria 9
Table 3: EMFASE success criteria categories 10
Table 4: Supporting criteria and parameters in relation to the MEM success constructs 13
Table 5: Framework scheme 16

List of figures

Figure 1: Method Evaluation Model 12
Figure 2: Empirical protocol to compare two SRA methods 17
Figure 3: Empirical studies timeline 20
Figure 4: Timeline for ongoing and planned empirical studies 21
Figure 5: Actual Effectiveness: Number of threats and security controls 22
Figure 6: Actual effectiveness 24
Figure 7: Quality of threats and security controls 25

Executive Summary

The main objective of WP1 of the EMFASE project is to develop a framework for empirical evaluation of methods for security risk assessment in the Air Traffic Management (ATM) domain. The purpose of the framework is to facilitate the evaluation and comparison of such methods and their techniques. The framework shall moreover aid stakeholders in selecting the most suitable method given the specific needs and available resources for conducting a security risk assessment.

In this document we present the refined EMFASE empirical framework. The deliverable shows the continued work after deliverable D1.2, which was the initial version of the framework. The framework is based on a Method Evaluation Model (MEM) for evaluating the success of a method, as well as a set of success criteria for security risk assessment methods in the ATM domain that we identified in collaboration with ATM security personnel. The EMFASE framework shall aid stakeholders in investigating which criteria actually contribute to the success of a security risk assessment method, and why.

More specifically, this document includes the following.

- A summary of the best practices for empirical methods that serve as guidelines for the EMFASE empirical experiments.

- A summary of the success criteria for ATM security risk assessment methods and how these are related to the MEM. The EMFASE framework and empirical studies shall help investigate and understand which criteria actually contributes to the success of security risk assessment methods and how. The criteria are classified into four main categories that are also used for structuring the empirical framework. These categories are method *process*, *presentation* of results, the actual risk assessment *results*, as well as *supporting material* for conducing security risk assessments.

- The EMFASE empirical framework that consists of two parts, namely a *framework scheme* and a *protocol* for conducting empirical experiments. The framework scheme is based on the identified success criteria and on the MEM. We also show how the experiments we have conducted so far are instantiated in the scheme. The protocol consists of two streams, namely an execution stream and a measurement stream. The former is the actual execution of the experiment where a security risk assessment method is applied, whereas the latter is the gathering of the data for the method evaluation.

- An overview of the EMFASE empirical studies conducted so far, as well as the evaluation and lessons learned regarding the EMFASE empirical framework.

1 Introduction

1.1 Purpose of the Document

The main objective of WP1 of the EMFASE project is to develop a framework for empirical evaluation of methods for security risk assessment in the Air Traffic Management (ATM) domain. The framework shall aid stakeholders in evaluating and comparing such methods, and in selecting the most suitable method given the specific needs and available resources for conducting a security risk assessment.

In this document we present the refined EMFASE empirical framework. The current framework is based on the initial version as presented in deliverable D1.2. The purpose of the document is to present the framework, as well as the evaluations and lessons learned from the empirical experiments conducted by the project during the last 12 months.

The empirical framework is based on a Method Evaluation Model (MEM) for evaluating the success of a method, as well as a set of success criteria for security risk assessment methods in the ATM domain. The success criteria were identified in collaboration with ATM security personnel. The EMFASE framework shall aid stakeholders in investigating which criteria actually contribute to the success of a security risk assessment method, and why.

The EMFASE empirical framework includes a scheme and a protocol for empirical studies. The scheme incorporates the MEM constructs and the success criteria, while the protocol describes steps that can be conducted for carrying out the empirical studies. The EMFASE empirical studies are based on existing practices and established empirical research methods.

More specifically the document is structured as follows. In Section 2 we briefly recapitulate on the best practices for empirical methods that guide the EMFASE empirical studies. In Section 3 we summarize the identified success criteria for ATM security risk assessment methods and explain how they are related to the MEM. In Section 4 we present the EMFASE empirical evaluation framework, including the scheme and the protocol. In Section 5 we give an overview of the EMFASE empirical studies that we have conducted so far. We also present the lessons learned regarding the further development of the EMFASE empirical framework. Finally we conclude in Section 6.

1.2 Intended Readership

The intended readers of this document 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 for all SESAR JU projects within the transversal areas of WP16 that are related to security management and risk assessment. For these stakeholders the document gives insight into some of the main criteria that should be fulfilled by methods for ATM security risk assessment, and also which 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. References to these projects are given in the relevant sections.

1.4 Glossary of Terms

Term	Definition
Control	Measure to modify or treat risk
Information security	Preservation of confidentiality, integrity and availability of information
Risk	The combination of the likelihood and consequence of an unwanted incident
Risk assessment	Overall process of risk identification, risk analysis and risk evaluation
Threat	Potential cause of an unwanted incident
Vulnerability	Weakness of an asset or a control that can be exploited by a threat

1.5 Acronyms and Terminology

Term	Definition
ANSP	Air Navigation Service Provider
ATM	Air Traffic Management
CLM	Concept Lifecycle Model
E-ATMS	European Air Traffic Management System
E-OCVM	European Operational Concept Validation Methodology
OFA	Operational Focus Area
MEM	Method Evaluation Model
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 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 [7].

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 [7] as outlined in Table 1. 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.

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

Table 1: Case study research process

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 [7][10]. 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 [4] 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 [7] for the process outlined in Table 1.

3 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 summarize the identified criteria and how they relate to the MEM. For further details we refer to D1.2.

3.1 Identified Success Criteria

Table 2 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 (approx. 30%). This indicates some spread in the opinions of the ATM stakeholders. Some of the less frequent statements 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.

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	
– Catalogue of threats and security controls	8
– Time effective	7
– Help to identify threats	6
– Applicable to different domains	5
– Common language	5
– Compliance	5
– Evolution support	5
– Holistic process	5
– Worked examples	5
Total	**159**

Table 2: Occurrences of reported success criteria

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 fulfil 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 [1].

- Compliance with ISO/IEC standards
- Well-defined terminology
- Documentation templates
- Modelling support
- Visualization
- Systematic listing
- Practical guidelines
- Assessment techniques
- Lists and repositories
- Comprehensibility of method outcomes

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.

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;

Table 3: EMFASE success criteria categories

3.2 Success Criteria and Risk Assessment MEM

The EMFASE empirical framework uses the Method Evaluation Model (MEM) proposed by Moody [6]. Methods have no 'implicit' value, only pragmatic value; a method in general, and a risk assessment method in particular, does not describe any external reality, so it cannot be true or false, but rather effective or ineffective.

The objective of EMFASE evaluation should therefore not be to demonstrate that the method is correct but that it is rational practice to adopt the method based on its pragmatic success. The *pragmatic success* of a method is defined as "the efficiency and effectiveness with which a method achieves its objectives" [6]. Methods are designed to improve performance of a task; efficiency improvement is achieved by reducing the effort required to complete the task, whereas effectiveness is improved by improving the quality of the result.

In addition to being efficient and effective, a method can be successful only if it is actually used in practice. The Technology Acceptance Model that is incorporated in the MEM captures this dimension by the constructs of *perceived ease of use*, *perceived usefulness* and *intention to use*.

Combining these aspects, Moody therefore argues that there are at least two dimensions of "success" that need to be considered, namely *actual efficacy* 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 1. 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

The arrows between the constructs in Figure 1 depict the hypothesized causal relationships between the constructs. For example, perceived usefulness is determined by actual effectiveness and perceived ease of use. EMFASE investigates these constructs and causal relationships to understand which features or properties of SRA methods that may contribute to them.

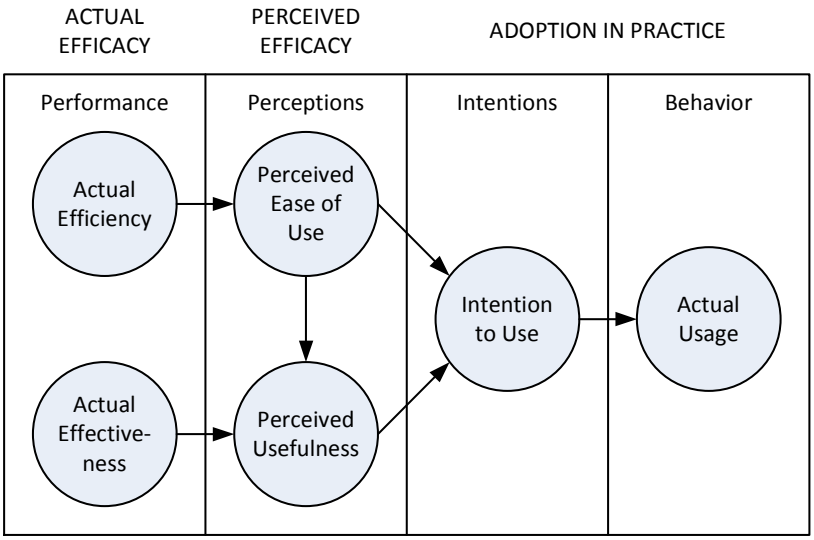

Figure 1: Method Evaluation Model

In Table 4 we give an overview of the relations between the identified criteria for the classification and evaluation of the security risk assessment methods and the MEM constructs we will evaluate during our experiments. A marked cell indicates that the supporting criterion/parameter may contribute to the fulfilment of the corresponding MEM constructs.

The EMFASE empirical studies are based on the identified success criteria, the MEM and the hypothesised relations between the criteria and the MEM constructs as presented in this section. In the next section we present our empirical framework for conducting such experiments.

Supporting Criteria	Success Constructs (MEM)			
	Perceived Ease of Use (PEOU)	Perceived Usefulness (PU)	Actual Efficacy (AE)	Intention to Use (ITU)
Clear steps in the process	X			
Specific controls		X	X	
Coverage of results		X		
Tool support		X		
Comparability of results	X			
Catalogue of threats and security controls	X	X	X	
Time effective	X			
Help to identify threats		X		
Applicable to different domains		X		
Well defined terminology	X			
Compliance with ISO/IEC standards		X		
Evolution support	X			
Holistic process		X		
Worked examples	X			
Documentation templates	X			
Visualization	X	X	X	
Systematic listing	X	X	X	
Modelling support	X			
Practical guidelines	X			
Assessment techniques		X	X	
Comprehensibility of method outcomes				X

Table 4: Supporting criteria and parameters in relation to the MEM success constructs

4 The EMFASE Framework

The objective of the framework is to support SESAR stakeholders in comparing two SRA methods and identify the preferred one with respect to the specific needs of the stakeholders for a specific security risk assessment. On the one hand the framework shall aid stakeholders in selecting the empirical studies or experiments that can be conducted in order to identify the preferred SRA method. On the other hand the framework is used by EMFASE to gather empirical data for providing guidance on which SRA methods or techniques to select given the stakeholder needs.

In the following we recall in Section 4.1 the purpose of the framework and who the main target group is. In Section 4.2 we present our empirical framework, consisting of a framework scheme and a protocol for conducting the experiments. We refer to D1.2 for a discussion of the relation to the security case of the ATM concept validation, which is relevant for both the E-OCVM [3] and the security reference material of SESAR project 16.06.02 [9].

4.1 Purpose and Target Group

The intended target group of the EMFASE framework is SESAR personnel that are responsible for developing the security case for the ATM concept validation. Such personnel are typically developers of Operational Focus Areas (OFAs) or developers of Operational Concepts. As such the EMFASE framework can support SESAR stakeholders in addressing ATM security and to conduct the security activities as specified by SESAR ATM Security Reference Material provided by project 16.06.02 [9].

The development and revision of the framework is based on the empirical studies conducted by the project, including (semi-) controlled experiments, complemented by surveys and literature studies. The framework is designed to enable the comparison of two given SRA methods so as to select the preferred method based on the stakeholders' needs, as well as the resources available to conduct the security risk assessment. The framework is therefore not developed to judge the absolute "goodness" of one SRA method, but rather how successful one SRA method is relative to another.

4.2 Empirical Framework

In the following we first present the scheme of the EMFASE empirical framework, which includes the success criteria and the related MEM constructs. Subsequently we present and explain the EMFASE protocol for conducting the experiments.

4.2.1 Framework Scheme

The scheme for the EMFASE empirical framework is shown in Table 5. In the following we explain its contents step by step.

The first column (#) refers to the EMFASE experiments that we have conducted or that are to come. The details of the experiment results are presented in deliverable D2.2 at M18 and D2.3 at M30. A brief overview of the experiments is given in Section 5, including some results from the two first experiments.

The second column (**type**) indicates whether or not the experiment is controlled (C). By "C-" we indicate that the experiment was only loosely controlled.

The **experiment context** describes characteristics of the experiment design. In our framework we have four such variables:

- Method experience: Indicates whether (Y) or not (N) the participants of the experiment have prior experience with the SRA methods object of study.

- Domain experience: Indicates whether (Y) or not (N) the participants of the experiment have experience from or background in the target system for the SRA.
- Model artefacts: Indicates whether the model artefacts, i.e. the documentation of risks and controls, are produced (Pd) by the participants during the experiments or provided (Pv) as part of the input material to the experiment.
- Time: Indicates whether the assigned/available time for the participants to complete the experiment tasks is varying (V) or fixed (F).

The **success variables** refer to the constructs of the MEM as shown in Figure 1 as well as to the identified SRA method success criteria. For each of the variables, experiments can be conducted to evaluate actual efficacy (A), perceived efficacy (P) or both (AP).

The **MEM** success variables are actual and perceived efficiency and effectiveness. For evaluating the actual effectiveness of an SRA method, experiments can be conducted in which the time is fixed. The actual effectiveness can then be evaluated by analysing the quality of the produced results. For evaluating the actual efficiency the quality is fixed instead. In that case, experiments are conducted to investigate the time that is required to conduct an SRA and reach a specific quality of results. The perceived effectiveness and efficiency can be investigated for both fixed and varying quality and time.

The remaining columns refer to the SRA success criteria presented in Section 3. As explained in that section we structured the success criteria by classifying them into four categories, namely process, presentation, results and supporting material. For each of the success criteria the framework and the scheme is a means to investigate whether it contributes to actual and/or perceived efficacy and to comprehensibility.

The **process** represents success criteria for the SRA process. The **presentation** concerns how the SRA results are presented and documented by using a given SRA. *Visualisation* refers to the suitability of the presentation format for specifying, analysing and understanding specific parts of an SRA, such as relations between specific threats, vulnerabilities and controls. It moreover refers to the suitability of the presentation for providing an overall view and understanding of the full results from an SRA for a given target of analysis. The *systematic listing* refers to the suitability of the presentation for listing, systematising, or sorting the SRA results, for example for information retrieval or categorisation. The *comprehensibility* refers to the extent to which risk documentation is understandable to end users and other stakeholders. The **supporting material** refers to any support that comes with an SRA, including tools, guidelines, work examples and catalogues of threats, vulnerabilities, controls, etc. In our scheme we have investigated catalogues, where *specific catalogues* are developed for a specific domain (such as ATM) and *generic catalogues* are domain independent. Note that we do not have a separate column for **results** since these are the method outcomes that are evaluated using the MEM constructs of efficiency and effectiveness, as well as comprehensibility.

#	Type	Experiment context				Success variables							
						MEM		Process	Presentation			Supporting material	
		Method experience	Domain experience	Model artefacts	Time	Efficient	Effective	Clear Process	Visualization	Systematic listing	Comprehensibility	Specific catalogue	Generic catalogue
1	C-	N	N	Pd	F	P	AP	P					P
2	C	N	N	Pd	F	P	AP	P				AP	AP
3	C	N	Y	Pd	F	P	AP	P				AP	AP
4	C	N	Y	Pd	F	P	AP	P					P
5	C	N	Y	Pv	F				AP	AP	AP		

Table 5: Framework scheme

The rows in Table 5 give an overview of the EMFASE experiments and how each of them is instantiated in the scheme. For cells that are unmarked the corresponding MEM variable or success criterion was irrelevant or not investigated. For further details about the experiments and the details the reader is referred to Section 5.

The participants of experiment 1 and 2 were MSc students, whereas the participants of experiment 3 and 4 were professionals. In all these experiments the time was fixed. Experiment 5 has currently been conducted with MSc students as participants. In the fifth experiment we investigate comprehensibility of risk documentation by comparing graphs and tables. The graphs and tables are risk model artefacts that in this experiment will be provided to the participants.

4.2.2 An Empirical Protocol to Compare Two SRA Methods

In this section we present an empirical protocol that can be applied to conduct empirical studies to compare two security risk assessment methods with respect framework scheme and to the success criteria identified in Section 3. This protocol was used in conducting the EMFASE experiments that have been completed so far, namely experiment 1 through 5.

Conceptually, the protocol is divided in two parallel streams that are merged in time as shown in Figure 2:

- The **execution stream** is the actual execution of the experiment in which the methods are applied and its results are produced and validated;
- The **measurement stream** gathers the quantitative and qualitative data that will be used to evaluate the methods.

Each stream is divided into three phases: *Training*, *Application* and *Evaluation*. We introduce each stream later in this section.

Three types of actors are necessary to execute the protocol (besides the researchers): *method designers*, *domain experts*, and *participants*. Method designers are the methods' inventors. Their main responsibility is to train participants in the method and to answer participants' questions during the Application phase. They evaluate group reports to determine if the method has been applied

correctly. Domain experts are usually industrial partners who introduce the application scenario to the participants. They evaluate the quality of the threats and security controls produced by each group of participants.

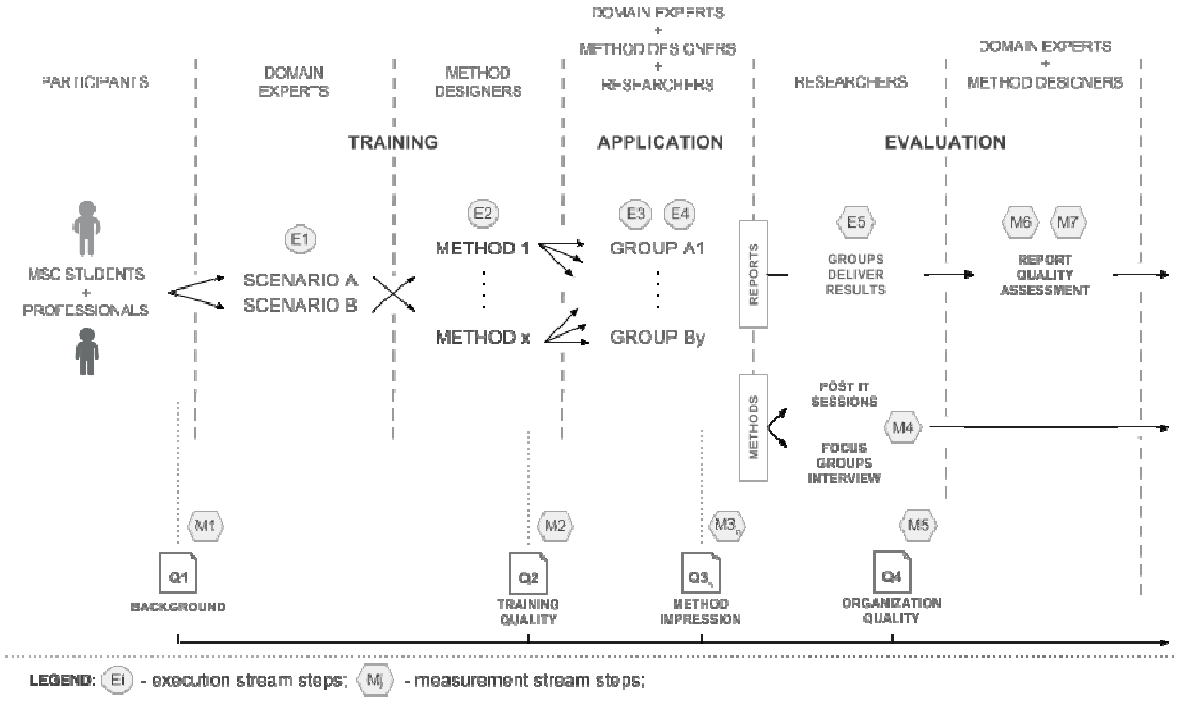

Figure 2: Empirical protocol to compare two SRA methods

The domain experts are also available during the Application phase to answer possible questions that the participants may raise during the SRA. Participants have to identify threats and security controls for an application scenario using the assigned method.

4.2.2.1 The Protocol's Execution Stream

Training. The goal of this phase is to train participants on the methods and the application scenarios.

- **E1** Participants attend lectures on the industrial application scenarios by the domain expert or by a trusted proxy.
- **E2** Participants attend lectures about the method by the method inventor or by a trusted proxy.

The first step 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. The rationale of the second step is to limit the threat to internal validity related to the implicit bias that might be introduced by having to train participant in one's own method as well as a competitor's method.

Application. The goal of this phase is to let the participants learn the method by applying it to the application scenario. The following two steps are therefore repeated at least a couple of times.

- **E3$_n$** Participants work in groups and apply the method to analyse the application scenarios.
- **E4$_n$** Groups give a short presentation about the preliminary results of the method application and receive feedback.

These steps address one of the major threats to internal validity, namely that the time spent in training participants was too short for them to be able to effectively apply the method. To mitigate this threat we have asked method designers and domain experts to be available to answer questions that participants may raise during the application of the methods. Further, step E3 should last at least two days of continuous work. The group presentation in E4 captures a phenomenon present in reality: meeting with customers in order to present progress and gather feedback. Participants may adjust their work along the received feedback. We do not consider this a bias because it is precisely what happens in reality. We considered the benefit for external validity greater than the threat to conclusion validity.

Evaluation. The goal of this phase is to collect the participants' results for evaluating the actual effectiveness of the methods.

- **E5** Groups deliver a presentation of the highlights and a final report documenting the application of the methods and the security analysis results.

4.2.2.2 The Protocol's Measurement Stream

Training. During this phase we capture the baseline knowledge of the participants (a possible confounding variable) and their initial understanding of the method (how easy/hard it *seems* to be).

- **M1** Participants are administered a questionnaire to collect information about their level of expertise in requirement engineering, security and on other methods they may know (Q1).
- **M2** Participants are distributed a post-training questionnaire to determine their initial perception of the methods and the quality of the tutorials (Q2).

The first step targets the threat to internal validity represented by participants' previous knowledge of the other methods. Collecting the background information about participants we control whether the participants have the same background and whether they have prior knowledge about methods under evaluation.

Application. The goal of this phase is to measure how the participants' perception of the methods changes the more they get acquainted with it.

- **M3$_n$** Participants are requested to answer a post-task questionnaire about their perception of the method (Q3$_n$) after each application session.

Evaluation. The goal of this phase is twofold. First, we *validate* whether the groups of participants have applied the method correctly and identified threats and security controls that are specific for the scenarios. Second, we *collect* the participant's perception and feedback on the methods through post-it note sessions and focus group interviews.

- **M4** Participants are divided in groups based on the assigned method. They are involved in focus group interviews where they are asked questions on their perception of the methods. A separate post-it note session is run with each group. In each session, the groups perform the following activities:
  - Post-it Notes. Each member of the group is requested to annotate on post-it notes 5 positive and 5 negative aspects of the applied method.
  - Post-it Notes Grouping and Prioritization. Each group has to hang the post-it notes on a wall and group notes that reports similar opinions about the aspects of the method. Once grouped, the post-it notes have to be listed in order of importance.
- **M5** Participants are requested to answer a post-task questionnaire about the quality of empirical study's organization (Q4).
- **M6** Method designers evaluate group reports. The method designers evaluate the quality of the method application. The level of quality is on a four item scale: *Unclear* (1), *Generic* (2), *Partial* (3) and *Total* (4).

- **M7** Domain experts evaluate group reports. The domain experts assess the quality of the threats and security controls. The level of quality scale is the same as in M6.

The last two steps address two issues that may affect both conclusion and construct validity. Any method can be *effective* if it does not need to deliver useful results for a third party (hence the evaluation by the domain expert). It can also be properly *easy to use* if participants do not follow it (hence the evaluation by the method designer).

5 Overview of EMFASE Empirical Studies

In this section we describe the empirical studies that we have conducted in EMFASE following the empirical protocol described in Section 4.2.2. As shown in Figure 3, we have conducted three types of empirical studies.

1. The first type of study aims to evaluate and compare textual and visual methods for security risk assessment with respect to their actual effectiveness in identifying threats, security controls, and in participants' perception. This type of study was conducted in Experiments 1 and 3.

2. The second type of study focuses on assessing the impact of using catalogues of threats and security controls on the actual effectiveness and perception of security risks assessment methods. This type of study was conducted in Experiments 2 and 4.

3. The third type of study aims to investigate the comprehensibility of risk models expressed in two modelling approaches: graphical and tabular. This type of evaluation was conducted in Experiment 5.

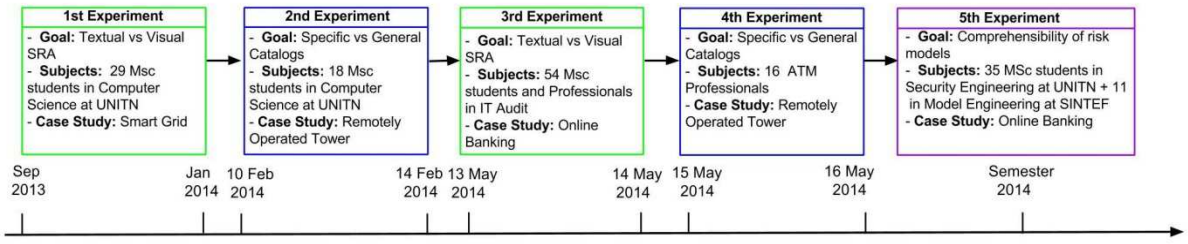

Figure 3: Empirical studies timeline

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).

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 [8] and CORAS [5]. 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.

We are preparing another series of experiments that will be carried out during 2015 and at the beginning of 2016. In these experiments, we will focus on comprehensibility (the third type of studies) from the perspective of security and ATM professionals. Figure 4 shows the timeline for the ongoing and planned experiments focusing on comprehensibility. More details on previous experiments and their preliminary evaluations are reported in D2.2.

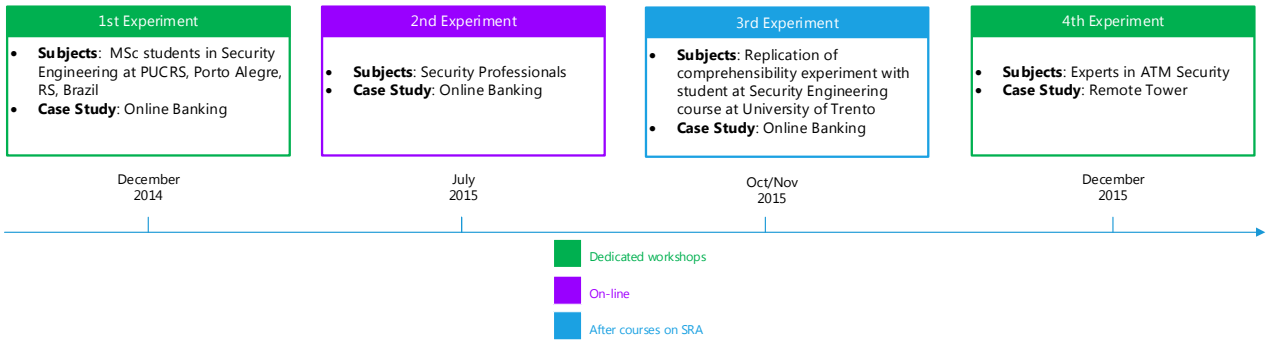

Figure 4: Timeline for ongoing and planned empirical studies

In addition to the above mentioned experiments we are also planning to have additional evaluation workshops with ATM experts and security professionals for the validation of the EMFASE empirical evaluation framework and of the EMFASE results and guidelines. We will participate both in SESAR Security Jamboree Meetings of WP16.6.2 to present and evaluate our findings, and we will also organize a Final Validation Workshop at the end of the project.

The Final Validation Workshop will be held during the period January/March 2016. The workshop will involve about 20 participants (from ANSPs, academy, and security consultancy). The validation method will be based on expert judgements, with an initial presentation of the EMFASE empirical evaluation framework and of the EMFASE outcomes to ATM and security professionals, as well as on a structured feedback collection through group discussions and questionnaires.

In what follows we provide an overview only of the empirical studies conducted with MSc students. For a more detailed report of all the conducted empirical studies we refer to EMFASE deliverables D2.2 and D2.3.

5.1 Evaluating and Comparing Visual and Textual Methods

The experiment involved 29 MSc students who applied both methods to an application scenario from the Smart Grid domain**.** CORAS [5] was selected as instance of a visual method, and EUROCONTROL SecRAM [2] as instance of a textual method.

5.1.1 Experimental Procedure

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.** Participants were given a 2 hours tutorial on the Smart Grid application scenario 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, and they were assigned to different security facets based on the experimental design.
- **Application.** 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. For each facet the participants:
  - Attended a two hours lecture on the threats and possible security controls specific to the facet, but not concretely applied to the scenario.
  - Had 2.5 weeks to apply the assigned methods 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**. In this 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 method. The Technology Acceptance Model (TAM) inspired the post-task questionnaires [6]. To prevent participants from "auto-pilot" answering, 15 out of 31 questions were given with the most positive response on the left and the most negative on the right. 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. The interview guide contained open questions about the overall opinion of the methods, whether the methods help in identification of threats and security controls and about the methods' possible advantages and disadvantages. The interview questions were the same for all the interviewees.

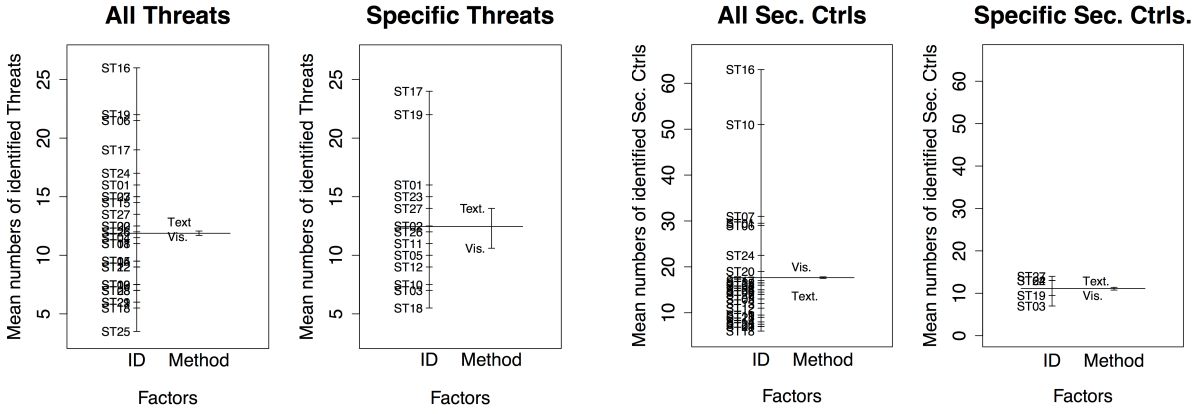

Figure 5: Actual Effectiveness: Number of threats and security controls

5.1.2 Experimental 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 specific threats and security controls.

Actual Effectiveness. Figure 5 (left) shows that the textual method is 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 5 (right) shows that the visual and textual method produce 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 there is no difference in the actual effectiveness of the visual and textual method for security risk assessment.

Participants' Perception. The average of responses shows that 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).

Perceived Ease of Use. The visual method is better than the textual with respect to overall Perceived Ease of Use 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.

Perceived Usefulness. The visual method is better than the textual with respect to Perceived Usefulness with statistical significance (Mann-Whitney test returns Z= -2.39, p-value = $1.7 * 10^{-2}$, es = 0.15).

Intention to Use. The visual method is better than the textual with respect to overall Intention to Use with statistical significance (Mann-Whitney test returns Z = -2.05, p-value = $3.9 * 10^{-2}$, es = 0.16).

Thus we can conclude that overall 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. Diagrams in visual method help participants in identifying threats and security controls because they give an overview of the assets and of possible threats agents and possible threat scenarios they initiate against the assets, while the identification of threats in the textual method is not facilitated by the use of tables. In fact, using tables makes it difficult to keep the link between assets and threats. Also, lower effectiveness and perception of the textual method can be explained by a poor worked example illustrating method application, and by the lack of software that supports the creation of the tables generated by the textual method.

5.2 Evaluating the Effect of Using Catalogues of Threats and Controls

The goal of this empirical study was to evaluate the effect of one of the success criteria that emerged from the focus group interviews with ATM professionals, namely the use a catalogue of threats and security controls. 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 [8]. The experiment involved 18 MSc students who were divided into 9 groups: half of them applied SESAR SecRAM with the domain-specific catalogues and the other half with the generic catalogues. Each group had to conduct a security risk assessment of the Remotely Operated Tower (ROT) operational concept.

5.2.1 Experimental Procedure

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

- **Training.** 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 the method tutorial and one of two sets of catalogues of threats and security controls. In addition, the participants of the groups that used the domain-specific catalogues signed a Non-Disclosure Agreement because the catalogues are confidential for EUROCONTROL. 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.** Once trained on the application scenario and the method, the participants had at least 6 hours in the class to reuse their security risk assessment with the help of catalogues. After the application phase participants delivered their final reports.
- **Evaluation.** 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.

5.2.2 Experimental 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 security controls propose real solution 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.

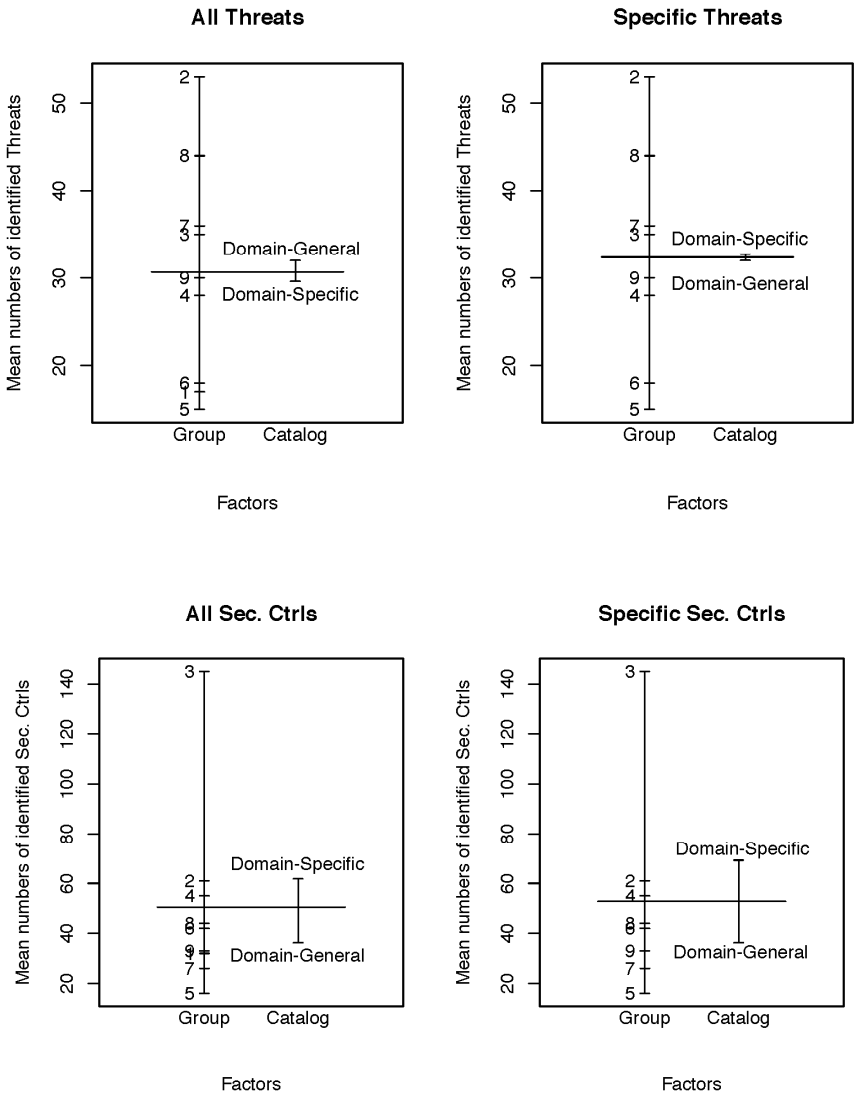

Figure 6: Actual effectiveness

Actual Effectiveness. First, we analysed the differences in the number of threats identified with each type of catalogue. As shown in Figure 6 (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.

We also compared the quality of threats identified with the two types of catalogues. Figure 7 (left) shows that the quality of threats identified with domain-specific catalogue is higher than the one of threats identified with domain-general catalogue. However, the Mann-Whitney test shows that the difference in the quality of identified threats is statistically significant only for specific threats (Z = −2.12, p-value = 0.046, r = −0.75).

Figure 6 (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). Figure 7 (right) shows that 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. This is not attested by the results of Mann-Whitney test for all security controls (Z = 0.77, p-value = 0.52, r = 0.26) and for specific security controls (Z = 0.31, p-value = 0.87, r = 0.11) show the difference in quality of security controls is not statistically significant.

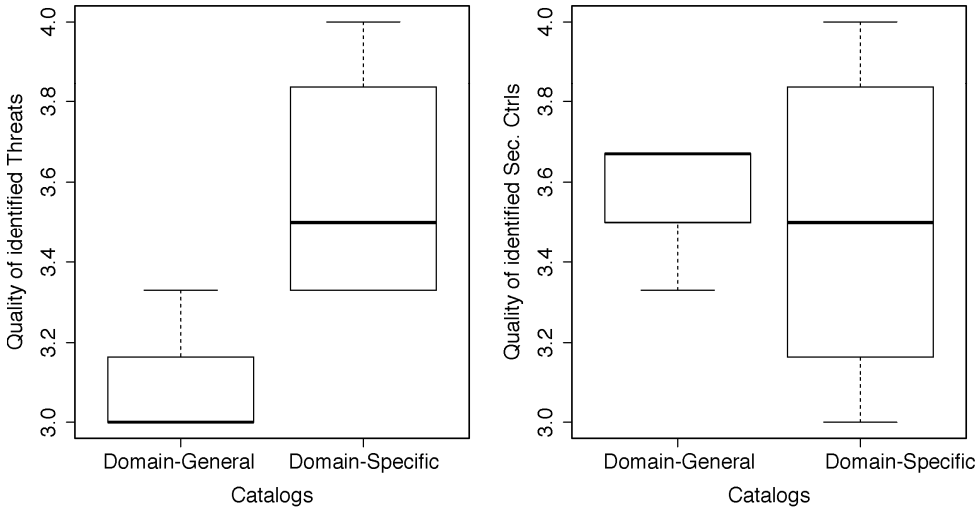

Figure 7: Quality of threats and security controls

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, p-value = $7 * 10^{-5}$, 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) 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.

Catalogues' Perception. The analysis of responses related to catalogues revealed no statistical significant difference between the types of catalogues overall perception, Perceived Ease of Use, Perceived Usefulness and Intention To Use. Only among good participants there is a 10% significant preference for domain-specific catalogues' Perceived Ease of Use but we cannot rely on this result because homogeneity of variance assumption is not met.

The results indicate that both types of catalogues have no significant effect on the 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. Thus, we can conclude that there is no difference in the actual effectiveness of the domain-specific and domain-generic catalogues. However, the overall perception and perceived usefulness of the method is higher when used with the domain-specific catalogues, which is considered easier to use than the domain-general one.

5.3 Evaluation and Lessons Learned

The results of the empirical studies reported in D2.2 and results of ongoing experiments will serve as a basis for a further redefinition the EMFASE Empirical Evaluation Framework. The main aspects of the empirical framework that could and should be updated are:

- the set of considered success criteria and their mutual relevance,
- the MEM constructs and their mapping with success criteria, and
- the experimental protocol adopted.

5.3.1 Success Criteria

The evaluation framework is based on a set of success criteria for SRA methods in the ATM domain. The results of the EMFASE empirical studies give insight into which criteria actually do have a significant effect on the success of a method. That is to say, the experiment results provide insights on how to complement and revise the set of success criteria, and thereby better adapt the framework to the criteria of significance and relevance for adoption.

Regarding the difference between “textual vs visual” methods, analysed in experiments 1 and 3, the results indicate that the textual methods have higher actual efficacy, since they do not require from the user to learn a new modelling notation, which may be difficult. Moreover textual methods do not require from the user to learn how to use a tool, which may be very time and effort consuming.

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

Regarding the difference between “domain specific vs domain generic catalogues”, analysed in experiments 2 and 4, the main findings show that, on average, there are no significant difference in actual efficacy of catalogues: security novices with catalogues performed the same as security experts without catalogues. This is really interesting and should be better interpreted and further discussed with domain experts. It also affects the evaluation criteria of learnability and memorability and their overall relevance in the empirical framework.

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

In general, catalogues 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 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 catalogues. Checklists and guidelines with detailed questions for each step of a security risk assessment method can be less prescriptive and better support the identification of uncommon and emerging threats and innovative controls.

Thus a very relevant criterion should be to have a methodology that supports creative and autonomous threats and controls identification through structured reasoning and checklist provisions more than a set of (supposed) exhaustive catalogues.

5.3.2 Mapping MEM Constructs to Success Criteria

The 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, as we plan to do, thanks to interactions with SESAR WP16.06.02 experts and during the Final Validation Workshop, as described in the introduction of Section 5.

For instance, the set of comprehensibility experiments have the main research questions slightly differently formulated as described in more details in deliverable D2.2 and in the referred EMFASE papers. This could also cause a rephrasing and different decomposition and/or mapping of the investigated SRA success criteria and MEM constructs.

Generally speaking, we can consider the comprehensibility as a specific attribute of the MEM 'actual effectiveness', while the effort and productivity are easily mapped on the MEM 'actual efficiency' attribute. The perceived ease of use is exactly used and measured in the same way as in the first empirical framework and related set of experiments in 2014-2015.

We should better investigate and explain this point to properly revise and adapt the MEM for EMFASE scope.

5.3.3 Review the Experimental Protocol

Our experiments have some limitations and "threats to validity" that should be solved in the future EMFASE round of experiments, mainly through minor modifications and customization for different situations of the proposed experimental protocol.

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

We would like to carry out 'on-line experiments' with also a small reimbursement for participants (that will be security professionals) to ensure a minimum set of 150 participant for experiment. Such as on-line experiment should be quite simple and short, thereby requiring the experimental protocol to be modified accordingly. Thus, there are still some remaining open issues such as "How long should an empirical study be?", "How to collect data?", and "How to overcome language gaps?"

There are several additional lessons related to the experiments' execution that we learned from our experience:

- **Group work**. The most basic lesson is that a group of minimum 2 participants should be considered as a basic unit for the experiments involving the application of full security risk assessment process. Because this process usually includes brainstorming activities that are difficult to perform individually. In some experiments the individual participants started complaining about it, as they wanted to work in groups.

- **Experimental Protocol.** At the beginning of the experiment we need to clearly present the flow of the experiment and the purpose of each questionnaire of form that we ask participants to fill in.

- **Scope of the analysis**. Another possible problem is that the scope or target of the analysis may be not clear for the participants. The tutorial on the application scenario should clearly state what is the target on which the participants need to focus their analysis.

- **Application Scenario**. In some experiments the participants complained about insufficient information about application scenario, as they were not experts in this topic. The possible solution in this case is the presence of the owner of the scenario in the class to answer questions from the groups.

- **Feedback**. It is good to participants to know if they correctly follow the steps of the assigned security risk assessment method. It can be done in form of step-by-step tutorial on the method with hands-on application of the steps and following wrap-up. The wrap-up sessions may be provided per group because some participants like students do not like public discussion of their results and prefer to ask individual questions. The final wrap-up should be followed by a refinement phase otherwise the feedback is useless. After finishing the analysis the groups should have additional 1-2 hours to finalize their reports.

The presented controlled experiments and the derived lessons learned have provided some potential solutions: 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. All these suggestions should be included as guidelines in the definition of a new protocol for experiment conduction in the final EMFASE evaluation framework.

Finally, 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. The SESAR JU develops guidance and support for building a security case. The EMFASE empirical framework should aid SESAR stakeholders in selecting the SRA methods that are best suitable for building and maintaining the security case.

6 Conclusion

In this document we have presented the refined EMFASE empirical evaluation framework, as well as the lessons learned from conducting the EMFASE experiments. The objective of the framework is to aid ATM security personnel and other relevant stakeholder in conducting empirical studies to compare security risk assessment (SRA) methods and identify the preferred one for a given need or task at hand. The comparison shall take into account both the particular stakeholder needs, as well as the resources available for conducting the security risk assessment.

The empirical framework has been developed to cover aspects of security risk assessment methods that have been identified as important for the ATM domain. The EMFASE framework classifies these aspects into four categories of success criteria for SRA methods in the ATM domain, namely process, presentation, results and supporting material. We identified the success criteria in collaboration with security personnel from the ATM domain. The empirical framework is used to investigate which of the success criteria actually contributes to the success of SRA methods, as well as how and why.

In addition to the framework scheme that makes use of the success criteria and the method success constructs of the Method Evaluation Model (MEM), the empirical framework comes with a protocol for conducting the empirical studies. The protocol consists of two streams, namely the execution stream and the measurement stream.

As discussed in this document, the EMFASE framework should aid ATM security stakeholders in developing the security case that is currently not supported by the E-OCVM. The framework should also contribute to the development of the security case as guided by the security reference material of SESAR project 16.06.02.

7 References

[1] EMFASE E.02.32: Selection of Risk Assessment Methods Object of Study, Deliverable D1.1, 2014

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

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

[4] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam and J. Rosenberg: Preliminary Guidelines for Empirical Research in Software Engineering. IEEE Transactions on Software Engineering, 28(8):721-734, 2002

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

[6] D. L. Moody: The Method Evaluation Model: A Theoretical Model for Validating Information Systems Design Models. In Proc. of the European Conference on Information Systems (ECIS'03), paper 79, 2003

[7] P. Runeson and M. Höst: Guidelines for Conducting and Reporting Case Study Research in Software Engineering. Empirical Software Engineering, 14:131-134, 2009

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

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

[10] R. K. Yin: Case Study Research: Design and Methods, SAGE Publications, 2003

-END OF DOCUMENT-