First Empirical Evaluation Framework and Existing Practices

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	First Empirical Evaluation Framework and Existing Practices
Deliverable ID	D1.2
Edition	00.01.02
Template Version	03.00.00
Task contributors	
SINTEF; University of Trento; Deep Blue	

Abstract

The main objective of EMFASE WP1 is to develop a framework for empirical evaluation of methods for security risk assessment in the ATM domain. This document presents the initial version of the framework. It is based on a method evaluation model (MEM) for evaluating the success of a method, as well as on a set of success criteria for security risk assessment methods for the ATM domain. The framework includes a scheme for conducting empirical studies that incorporates the MEM constructs and the success criteria, as well as the EMFASE protocol for conducting empirical studies to compare security risk assessment methods.

Authoring & Approval

Prepared By - *Authors of the document.*		
Name & Company	Position & Title	Date
Bjørnar Solhaug (SINTEF)	WP1 leader	23/06/2014
Federica Paci (UNITN)	Project member	02/08/2014
Alessandra Tedeschi (DBL)	Project member	31/07/2014
Kate Labunets (UNITN)	Project member	02/08/2014
Martina de Gramatica (UNITN)	Project member	31/07/2014

Reviewed By - *Reviewers internal to the project.*		
Name & Company	Position & Title	Date
Elisa Chiarani (UNITN)	Project manager	26/08/2014
Federica Paci (UNITN)	Project member	27/08/2014

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	Coordinator	08/09/2014

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	23/06/2014	Document structure	B. Solhaug	Document creation
00.00.02	31/07/2014	Working document	A. Tedeschi, M. de Gramatica	Section 3
00.00.03	02/08/2014	Working document	F. Paci, K. Labunets	Section 4.3.2 and Section 5
00.00.04	07/08/2014	Working document	B. Solhaug	First draft of full document
00.01.00	12/08/2014	Working document	B. Solhaug, A. Tedeschi	Finalized for internal review
00.01.01	27/08/2014	Working document	E. Chiarani, F. Paci	Review and quality check

00.01.02	03/09/2014	Working document	B. Solhaug, F. Paci	Revision after review
00.01.02	03/09/2014	Final draft	F. Massacci	Approval for submission to PO to get feedback

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 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 our initial version of the EMFASE empirical framework. The framework is based on a Method Evaluation Model (MEM) for evaluating the success of a method. It is moreover based on 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.

More specifically, this document makes the following contributions.

- An overview of state of the art and best practices for empirical methods, including guidelines for how to conduct empirical studies.

- Success criteria for ATM security risk assessment methods and how these are related to the MEM. The success criteria were identified in collaboration with ATM security personnel. 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 initial and preliminary 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, including the reporting of some of the results.

The first EMFASE empirical framework as presented in this document is based on the results thus far in the project. The success criteria, how they are related to the MEM and the empirical framework will be revised and elaborated during the course of the project as we gather more empirical data and a better understanding of which criteria actually contribute to the quality of security risk assessment methods.

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 our initial version of the EMFASE empirical framework. The framework is based on a Method Evaluation Model (MEM) for evaluating the success of a method. It is moreover based on 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 initial 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 give a brief overview of the state of the art and best practices within empirical methods. In Section 3 we present the identified success criteria for ATM security risk assessment methods, introduce the MEM, and relate the success criteria to the MEM. In Section 4 we present the first EMFASE empirical evaluation framework, including the scheme and the protocol. The section also relates the EMFASE framework to the concept validation and the security case of E-OCVM, and to the SESAR security reference material of project 16.06.02. In Section 5 we give an overview of the EMFASE empirical studies that we have conducted so far and report on some of the results. 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 Empirical Methods – State of the Art and Best Practices

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 [12].

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. In this section we give a brief overview of relevant research methods with reference to literature, and we describe the guidelines that we follow. Note that we do not include action research in this overview as this method is not relevant for the EMFASE objectives.

2.1 Overview of Existing Empirical Methods

There are various kinds of methods that can be used for conducting empirical research. The methods vary on realism, precision and generality [8], and therefore serve different purposes in empirical studies. EMFASE combines several of such methods in the conducted studies, the most important of which are experiments, case studies and surveys.

- A (controlled) experiment involves the measuring of the effects of manipulating variables where the subjects (participants) are assigned to treatments by random [11][12][17]. Experiments score high on precision, but are typically weaker on realism.
- A case study is an empirical method that investigates a contemporary phenomenon in its context [1][11][12][18]. While case studies score high on realism, they are typically characterized by lack of experimental control [1] and that they may score low on precision.
- A survey involves the collection of standardized information from a specific (sample) population, usually by means of interviews or questionnaires [11][12]. Surveys can be high on generality, but may score lower on realism. The level of precision depends on the kind of survey that is conducted as it can vary from unstructured interviews to highly structured questionnaires.

The EMFASE case studies will mainly be conducted by observations [12] in order to investigate how specific assessment tasks and activities are conducted by ATM practitioners. The gathered data may be complemented by interviews or questionnaires. In addition to the experiments, observations and questionnaires, EMFASE makes use of literature reviews and expert judgments. Literature reviews include the investigation of any existing theories on method evaluation, as well as historical or archival data on security risk assessments that have been conducted within the ATM domain. Expert judgments are important for evaluating the quality of the results that are produced by the subjects of the controlled experiments.

The purposes of the three mentioned empirical research methods can be distinguished by the following classification [11][12].

- Exploratory: Finding out what is happening, seeking new insights and generating ideas and hypotheses for new research.
- Descriptive: Portraying a situation or phenomenon.
- Explanatory: Seeking an explanation of a situation or a problem, mostly but not necessary in the form of a causal relationship.

The research process can be characterized as fixed or flexible. In a fixed process, all parameters are defined at the initial phase of the study, whereas in a flexible process, the parameters of the study may be changed during the course of the project.

The characteristics of the methods for empirical research can be summarized by the overview of Table 1, adapted from [12]. Notice that that the given characteristics are only the ones that are considered as the primary for the method in question. For example, while EMFASE experiments for explanatory purposes, also the questionnaires, interviews and observations may serve this objective.

Method	Primary objective	Primary data	Design
Survey	Descriptive	Quantitative	Fixed
Case study	Exploratory	Qualitative	Flexible
Experiment	Explanatory	Quantitative	Fixed

Table 1: Characteristics of empirical research methods

In addition to method triangulation, i.e. the investigation of a topic with different empirical methods, EMFASE will as much as possible and when adequate use data triangulation, observer triangulation and theory triangulation. Data triangulation is the use of more than one data source, observer triangulation is the use of more than one observer, and theory triangulation is the use of alternative theories or viewpoints [12][15].

Case studies are an established research method is areas like psychology, sociology and political science [18], but are increasingly used also within information systems [1] and software engineering [6][12]. Software engineering involves the development, operation and maintenance of software and related artefacts. In [12] it is argued that research on software engineering to a large extent aims at investigating how these activities are conducted by software engineers and other stakeholders under different conditions. EMFASE is not concerned with engineering, although the development of ATM operational concepts is an important part of the study context. EMFASE is rather concerned with SRA and the methods for conducting the assessments. More precisely, EMFASE is investigating how security risk assessment is conducted by analysts, ATM practitioners and other stakeholders under different conditions, as well as how the results are used.

As mentioned above, we seek in this project not only to understand which SRA methods work under which conditions, but also *what* makes them work. Seeking the explanation of how and why SRA methods work in the context of ATM security is part of the theory building of Work Package 3. In [1] it is argued that empirical studies are particularly appropriate for problems in which theory is in its formative stage. This is indeed the case for the efficiency and effectiveness of SRA methods in the complex ATM domain where the security case of the validation of operational concepts is still not adopted by the E-OCVM [4]. The same paper moreover highlights the usefulness of case studies for "practice-based problems where the experiences of the actors are important and the context of action is critical" [1]. This clearly fits with the goal of EMFASE, namely to provide the relevant stakeholders with the means to select the SRA methods that are best suited for the task at hand.

2.2 Guidelines for Empirical Studies

EMFASE follows established guidelines and best practices for how to conduct and report empirical studies. In the following we give a high-level description of the process we follow, and we highlight some requirements that should be fulfilled in such studies.

For case study research there are several instruction books available from social sciences [11][15][18] that have been used also within information systems research [1] and software engineering research [12]. Guidelines and handbooks on empirical research targeting software engineering in particular

have also started to emerge [6][17]. Based on such guidelines we follow the research process of [12] as outlined in Table 2. 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	Case study design: Objectives are defined and the case 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 studied case
4	Analysis of collected data
5	Reporting

Table 2: Case study research process

Step 1 involves defining the objectives of the case study, i.e. what to achieve and which research questions to investigate. The case, i.e. what to be studied, must also be specified. For EMFASE, the case 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 case 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 [12][18]. 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 [6] 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 [12] for the process outlined in Table 2.

Most of the aforementioned handbooks and guidelines on empirical research are on empirical studies in general, or adapted to information systems or software engineering. Because EMFASE is concerned with SRA, the objective is to evaluate the value of SRA methods. In such evaluations we make use of the Method Evaluation Model (MEM) introduced by Moody [9] that we introduce in the next section.

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 need to identify the criteria with respect to which the methods shall be evaluated. There are of course many different parameters and aspects that can be considered for the classification and evaluation of methods for security risk assessment. In the EMFASE project, we derived the success criteria in close collaboration with ATM security stakeholders. In this section we present the success criteria identification process and the identified criteria, before introducing the Method Evaluation Model. Finally we relate the success criteria to the MEM by describing the hypothesized relations between each criterion and the constructs of the MEM.

3.1 Success Criteria Identification Process

We carried out an initial survey among ATM stakeholders to success criteria for SRA methods during the 6th Jamboree of the SESAR project 16.06.02, held in Brussels on 12 November 2013. All the raw data collected during the survey were analysed through coding techniques drawn from grounded theory [16]. After this analysis a first set of high-level success criteria was identified. They were reviewed, categorized and complemented by security experts in the EMFASE consortium and their first set was presented in D1.1 [2].

Subsequently we started to further analyse the identified success criteria in order to relate them to the Method Evaluation Model (MEM) [9]. Our hypothesis is that if a criterion really improves the success of an SRA method, it can be related to one or more of the constructs of the MEM. In this section we will show a first set of hypothetical links between the success criteria and the MEM, and explain how the EMFASE framework will be developed to investigate these links. The causal explanations will be provided later by WP3 based on the empirical frameworks and the results of the EMFASE empirical studies.

In order to properly observe, collect evidences and assess the identified criteria, they should be carefully decomposed in measurable indicators that closely depend on the specific experimental setting. The experiment hypotheses and the experimental protocol should be carefully designed for each evaluation experiment in the EMFASE empirical framework.

We preliminarily categorized the identified success criteria 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 stated above, our initial hypothesis is that each success criterion contributes to one or more of the MEM constructs, i.e. that the fulfilment of the success criteria contributes to the success of a security risk assessment method. The hypothesis will be investigated in the project, and the results will be used to develop the EMFASE framework, to define the guidelines for security risk assessment method selection in ATM, and to derive the causal explanations. The guidelines will be delivered by Work Package 1 at M24, whereas the development of the causal explanations is the task of Work Package 3. Our initial EMFASE framework is presented in Section 4 in this deliverable.

The success criteria and their relationship with MEM constructs will be further investigated and validated in a set of semi-structured interviews with security experts (not only from the ATM domain) during the autumn of 2014.

Figure 1 summarises the success criteria identification process carried out during the EMFASE first year of activity. In the continuation of the project we will revise the identified criteria, their categorization, and their relations to the MEM constructs based on new insight and the knowledge gathered during EMFASE experiments.

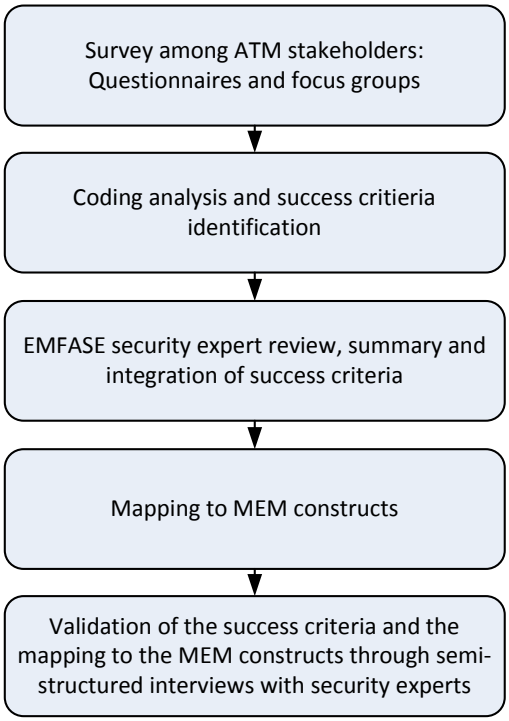

Figure 1: EMFASE criteria identification process

3.2 Coding of Survey Results

The survey included a questionnaire that was filled in by the participants individually, as well as group interviews where the participants were organized into separate focus groups of 5-6 people in each group.

The participants were all professionals from different organizations and enterprises within the aviation domain. While their background in security and risk management was of varying degree, they were all to some extent required to consider security risks and their mitigation as part of their work. The participants were hence a representative selection of ATM stakeholders with qualified opinions about and insights into the methodical needs for conducting a security risk assessment. The questionnaire included an open question about the main success criteria for security risk assessment methods, and this topic was also covered by the interviews.

We analysed the questionnaire answers and the interview transcripts using *coding* which is a content analysis technique that allows extracting qualitative data to be analysed quantitatively [16]. 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.

Raw textual data, interviews transcripts in this case, are disassembled and assigned to different pre-defined codes. According to the literature, code is a word or a short phrase that symbolically summarizes, condense and reduce the meaning of the data itself through an evocative meaning. To a higher level of analysis, codes can be clustered in more abstract categories that are used to generate

a theory (grounded theory).The frequency of the occurrence of each code determines its salience, namely the importance of a code in a text, while correlating codes with each other (co-occurrence) is a way to understand how topics under debate are framed.

The coding process can be done manually, which can be as simple as highlighting different concepts with different colours, or fed into a software package. Qualitative software packages include, for example Atlas.ti, QDA Miner and NVivo. These programs do not supplant the interpretive nature of coding but rather are aimed at enhancing the analyst's efficiency at data storage/retrieval and at applying the codes to the data. Many programs offer efficiencies in editing and revising coding, which allow for work sharing, peer review, and recursive examination of data.

When coding is complete, the analyst prepares reports via a mix of summarizing the prevalence of codes, discussing similarities and differences in related codes across distinct original sources/contexts, and comparing the relationship between one or more codes.

The EMFASE coding analysis of survey with ATM stakeholders was conducted as follows.

1. We analysed the responses to the open question and the interview transcripts to identify the recurrent patterns (codes) about the success criteria for the security risk assessment methods. We used the Atlas.ti software package.
2. The identified codes were grouped by their similarity and classified into categories.
3. For each category we counted the number of statements as a measure of their relative importance.
4. We employed multiple coders working independently on the same data (typically two or even three) and then compared the results. This minimizes the chance of errors from coding and increases the reliability of results.

3.3 Identified Success Criteria

Table 3 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.

The success criteria will be continuously validated and reviewed during the project lifecycle by expert judgment and by additional findings obtained during experiments.

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 3: Occurrences of reported success criteria

In our classification and evaluation of security risk assessment methods we will 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. This is an initial set of parameters, a more detailed description of which is presented in D1.1 [2], that is likely to be extended and/or revised during the course of the EMFASE project:

- 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, namely *process*, *presentation*, *results*, and *supporting material*. These four categories are in turn used for structuring the EMFASE empirical framework. As a result the following classification is the initial and preliminary scheme for supporting the method evaluation in EMFASE; in the continuation of the project we will revise this scheme based on new insight, knowledge or further elements for extracting relevant criteria.

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

3.4 Success Criteria and Risk Assessment Methods Evaluation Model

The EMFASE empirical framework uses the Method Evaluation Model (MEM) proposed by Moody [9]. 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" [9]. 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 2. 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 2 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.

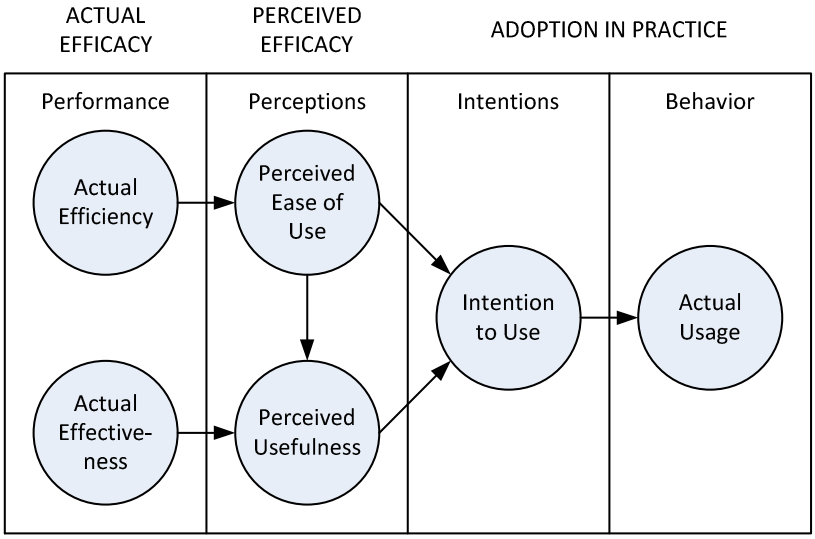

Figure 2: Method Evaluation Model

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 initial 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 explain in Section 4.1 the purpose of the framework and who the main target group is. In Section 4.2 we relate the EMFASE framework to the security case of the ATM concept validation, which is relevant for both the E-OCVM [4] and the security reference material of SESAR project 16.06.02 [14]. In Section 4.3 we present our initial empirical framework, consisting of a framework scheme and a protocol for conducting the experiments.

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 [14].

The current framework is the initial version, and it will be revised and further developed until M24 of the EMFASE project. The developments and revisions will be based on the empirical studies conducted by the project, including (semi-) controlled experiments and case studies (observations), 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.

In addition to the EMFASE empirical framework, the project will at M24 deliver a set of guidelines to aid stakeholders in selecting the SRA method or techniques that are suitable for specific needs. The guidelines will be based on the results of all of the empirical studies of the project over the two first years, as well as on the causal explanations that will be developed within Work Package 3. The target group of the EMFASE guidelines includes also Air Navigation Service Providers (ANSP) that should be able to use the guidelines to rate the suitability or success of an SRA method given the needs of the stakeholders. The guidelines will be presented as part of deliverable D1.3.

4.2 The Security Case of the E-OCVM

Before introducing the initial version of the EMFASE empirical framework we describe here briefly how it relates to the security case of SESAR 16.06.02 and to the E-OCVM [4]. The Concept Lifecycle Model (CLM) of the E-OCVM is depicted in Figure 3. The figure includes the validation phases the initial ATM needs (V0) to the eventual decommissioning (V7). The scope of the E-OCVM includes the phases V1-V3 as shown in the figure.

In its current version, the E-OCVM does not include security engineering activities such as security requirements engineering or security architecture; to indicate its relevance we added it to the bottom of Figure 3. The E-OCVM specification states that the security case, which we have added to the top of the figure, "may and should be developed", but it is still not part of the concept validation.

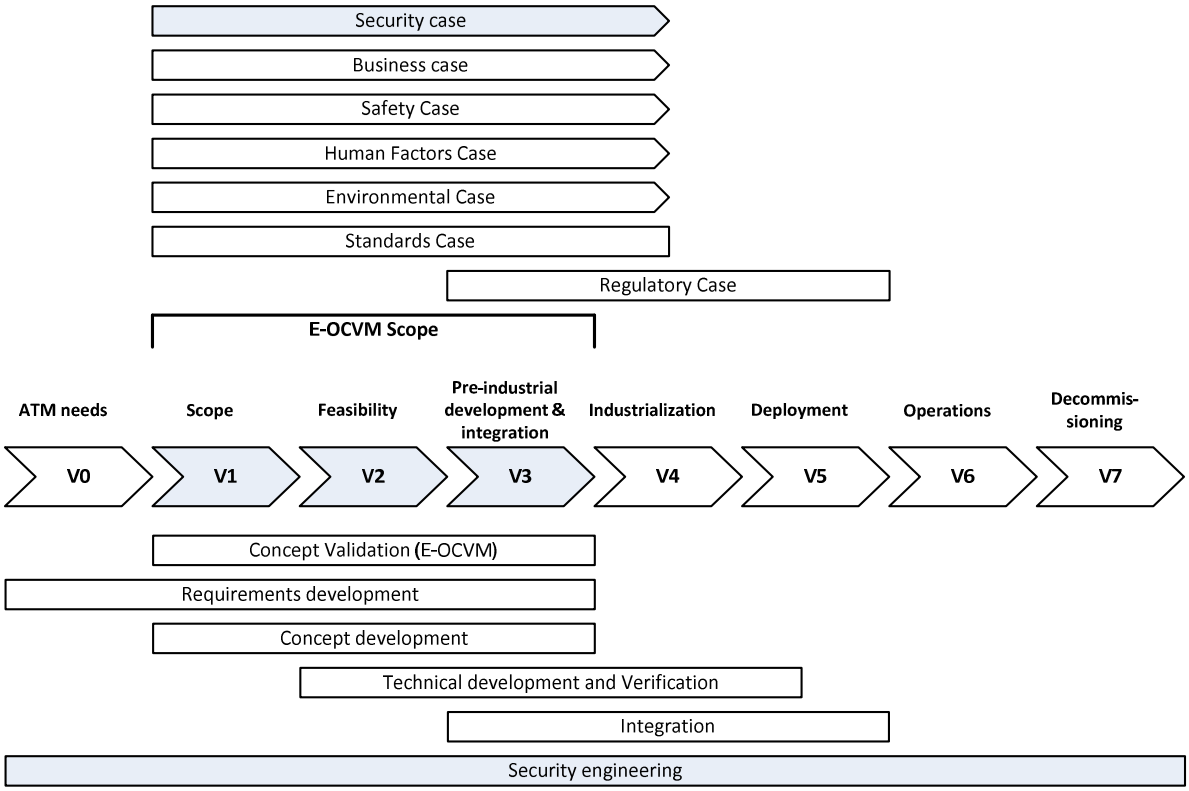

Figure 3: Adding the Security Case to the E-OCVM

The Security Reference Material of SESAR 16.06.02 provides guidance on the security activities that OFAs shall perform to build such a security case and move on to the deployment phase. The security activities include conducting security risk assessments and identifying adequate security controls for unacceptable risks.

For the ATM security personnel to effectively and efficiently conduct the security risk assessments the security reference material and the E-OCVM should include guidance on which SRA methods to use. The EMFASE project has the potential to support the development of such guidance by the identification of the SRA techniques and supporting material that are adequate for building the security case. The EMFASE framework should moreover support ATM stakeholders in conducting their own empirical studies in order to select the SRA methods that fulfil the needs in validating security of operational concepts.

4.3 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.

4.3.1 Framework Scheme

The scheme for the initial 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 will be presented in deliverable D2.2 at M18. A brief overview of

the experiments is given in Section 5, including some results from the two first experiments. The fifth experiment is upcoming, but is currently being designed.

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 initial 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 2 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 is currently being designed and will be conducted during the autumn of 2014 with MSc students as participants. In this experiment we will 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.

4.3.2 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 4.

Conceptually, the protocol is divided in two parallel streams that are merged in time as shown in Figure 4:

- 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.

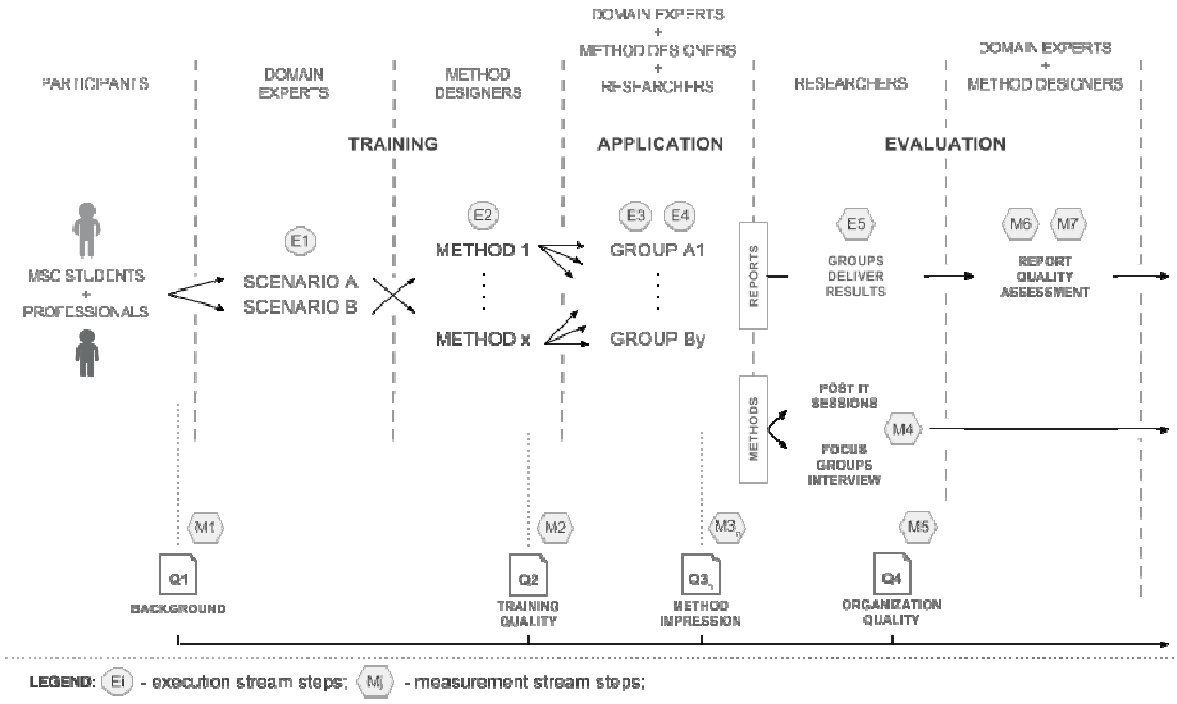

Figure 4: Empirical protocol to compare two SRA methods

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.

4.3.2.1 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.

4.3.2.2 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 4.3.2. As shown in Figure 5, we have conducted two types of empirical studies. The first type aims to evaluate and compare textual and visual methods for security risk assessment with respect to their actual effectiveness in identifying threats and security controls and participants' perception. The second type of studies focuses on assessing the impact of using catalogues of threats and security controls on the actual effectiveness and perception of security risks assessment methods. Both type of studies have been first conducted with MSc students and then with professionals. In what follows we provide an overview only of the empirical studies conducted with MSc students. We will provide a detailed description of all the conducted empirical studies in EMFASE deliverable D2.2 – First Evaluation Report.

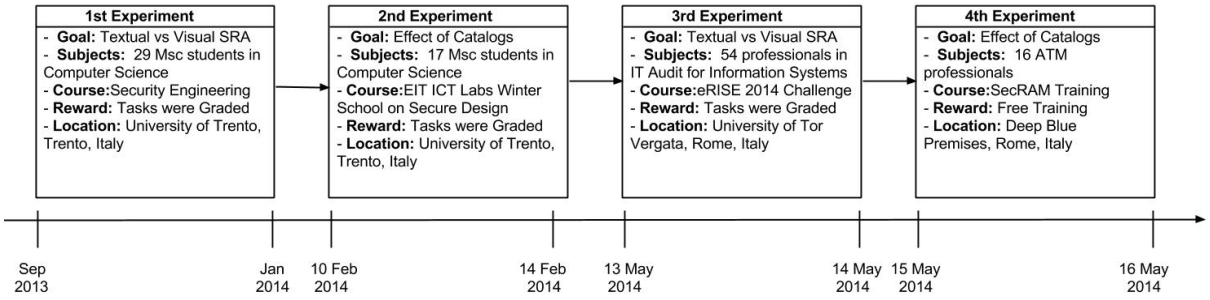

Figure 5: Empirical studies timeline

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 [7] was selected as instance of a visual method, and EUROCONTROL SecRAM [3] 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 [9]. 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.

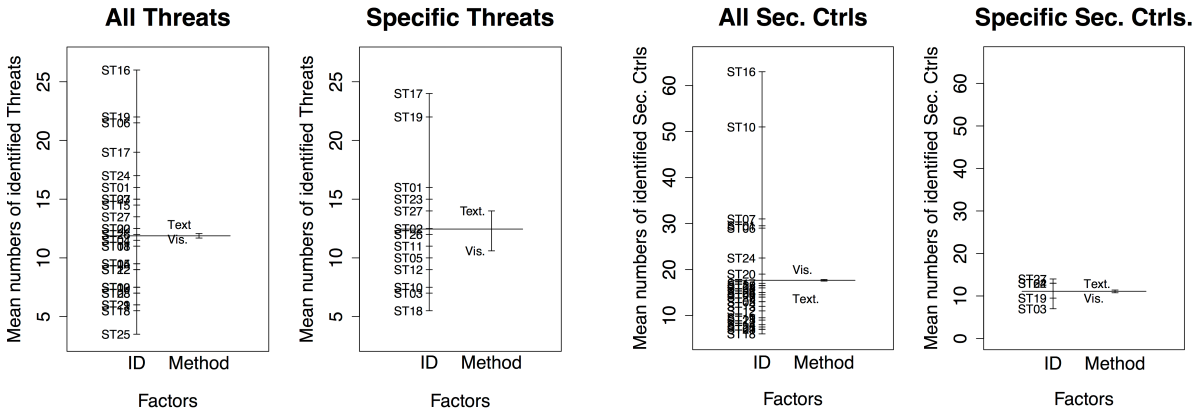

Figure 6: Actual Effectiveness: Number of threats and security controls

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 6 (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 6 (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 [13]. 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.

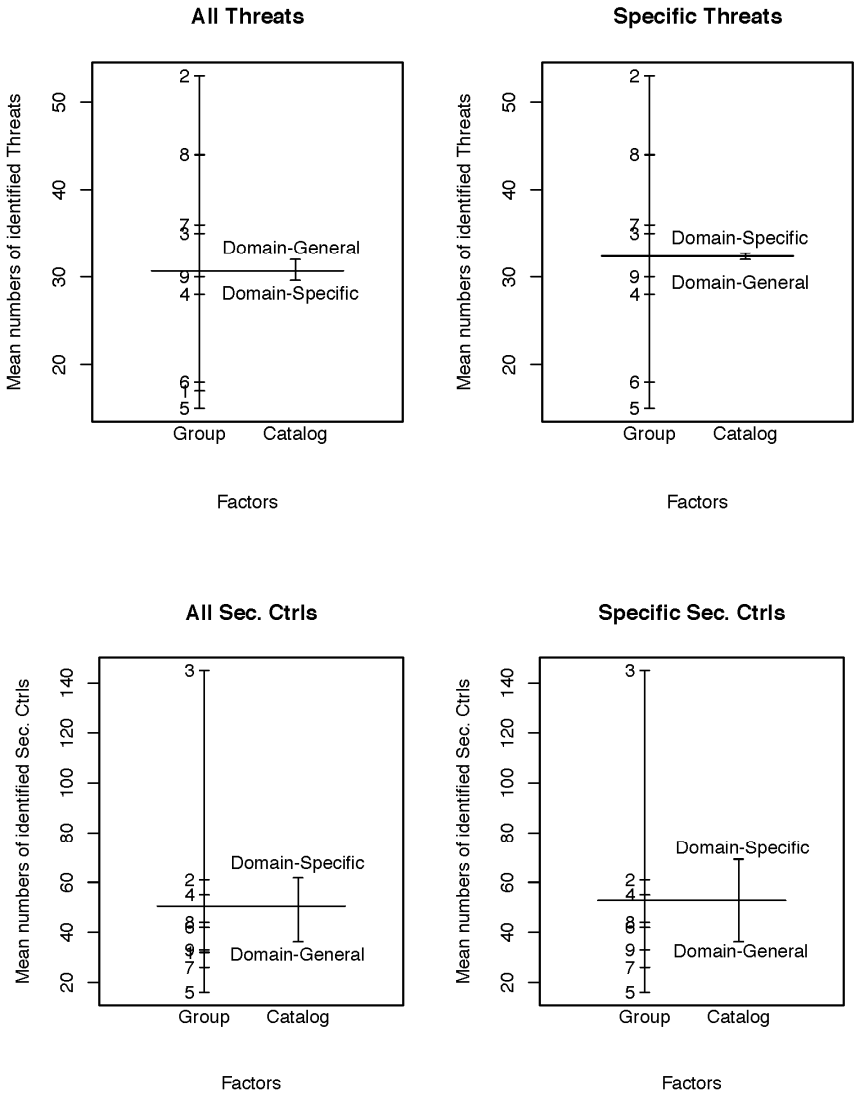

Figure 7: Actual effectiveness

Actual Effectiveness. First, we analysed the differences in the number of threats identified with each type of catalogue. As shown in Figure 7 (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 8 (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 7 (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 8 (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.

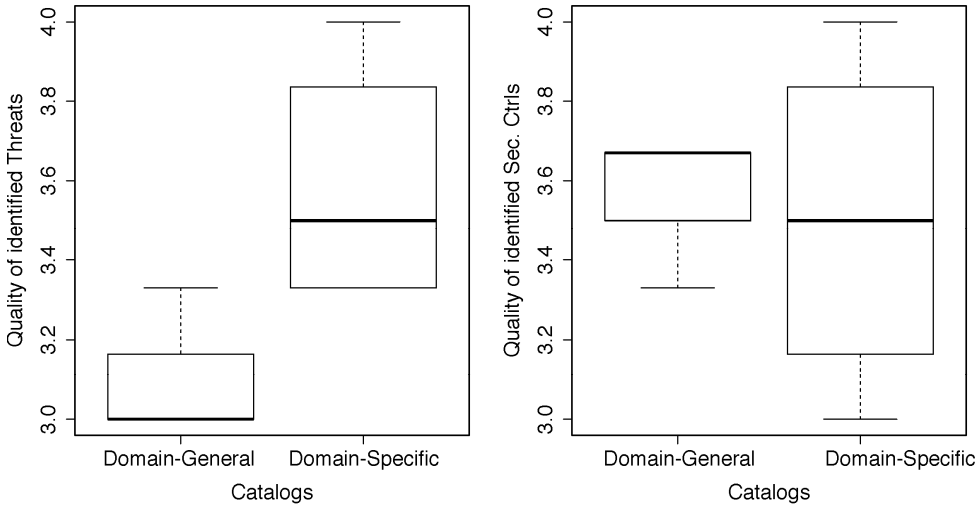

Figure 8: Quality of threats and security controls

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.

6 Conclusion

In this document we have presented the first EMFASE empirical evaluation framework. 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. During the course of the project the set of success criteria, and thereby also the empirical framework, will be revised and further elaborated.

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.

EMFASE will moreover deliver a set of guidelines for selecting SRA methods or techniques for the ATM domain. While the framework can be used by stakeholders to do comparison of SRA methods, the guidelines will be based on the empirical findings of the EMFASE project. The findings will be used in the context of EMFASE Work Package 3 to identify the causal relationship between SRA success criteria and the success constructs of the MEM.

The empirical framework presented in this document is the initial and preliminary EMFASE empirical framework. The framework will be further developed during the course of the project, and a revised version will be provided in deliverable D1.3 at M24. 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] I. Benbasat, D. K. Goldstein and M. Mead: The Case Research Strategy in Studies of Information Systems. MIS Quarterly, 11(3):369-386, 1987

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

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

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

[5] B. Flyvbjerg: Five Misunderstandings about Case-Study Research. Qualitative Inquiry, 12(2):219-245, 2006

[6] 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

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

[8] J. E. McGrath: Groups: Interaction and Performance, Prentice Hall, 1984

[9] 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

[10] N. Rescher: Methodological Pragmatism: Systems-Theoretic Approach to the Theory of Knowledge, Basil Blackwell, Oxford, 1977

[11] C. Robson: Real World Research, 2nd edition, Blackwell Publishing, 2002

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

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

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

[15] R. E. Stake: The Art of Case Study Research, Sage, 1995

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

[17] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell and A. Weslén: Experimentation in Software Engineering, Springer, 2012

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

-END OF DOCUMENT-