A First Empirical Evaluation Framework for Security Risk Assessment Methods in the ATM Domain

K. Labunets[1], F. Massacci[1], F. Paci[1], M. Ragosta[2] B. Solhaug[3], K. Stølen[3], A. Tedeschi[2]

[1]University of Trento
Via Sommarive 9
38122 Trento - Italy
{name.surname}@unitn.it

[2]DeepBlue Srl
Piazza Buenos Aires 20,
00198 Roma - Italy
{name.surname}@dblue.it

[3]SINTEF ICT
Forskningsveien 1a,
0314 Oslo - Norway
{name.surname}@sintef.no

***Abstract*— Evaluation and validation methodologies are integral parts of Air Traffic Management (ATM). They are well understood for safety, environmental and other business cases for which operational validation guidelines exist which are well defined and widely used. In contrast, there are no accepted methods to evaluate and compare the effectiveness of risk assessment practices for security. The EMFASE project aims to address this gap by providing an innovative framework to compare and evaluate in a qualitative and quantitative manner risk assessment methods for security in ATM. This paper presents the initial version of the framework and the results of the experiments we conducted to compare and assess security risk assessment methods in ATM. The results indicate that participants better perceive graphical methods for security risk assessment. In addition, the use of domain-specific catalogues of threats and security controls seems to have a significant effect on the perceived usefulness of the methods.**

Keywords:Empirical study, controlled experiment, security risk assessment methods, method evaluation.

I. INTRODUCTION

The Single European Sky ATM Research (SESAR) aims at developing the future European ATM System and associated operational and technological improvements. In this multi-stakeholder context ATM security is a key enabler to ensure the overall performance of the ATM System. At the same time, however, ATM security can be seen as a significant source of costs whose return on investment is not fully justified or evaluated. The trade-off between security and cost can only be evaluated by risk analysis. Validation methodologies are well understood and widely deployed in ATM for a wide number of aspects, ranging from safety to environment, but not for security. Indeed, the effectiveness of security risk assessment practices, as well as the comparative evaluation of such practices, is largely uncharted territory. It is unclear to what degree these practices and their activities provide security and whether or not they give return on investment. Furthermore, there are no accepted methods to evaluate or compare security practices and to decide that activity X works better than activity Y in a given setting. A central question is thus: How can SESAR stakeholders know that their methods for ensuring security in the complex ATM domain really work? Would additional expensive security analysis and measures be worth the cost? The SESAR WP-E EMFASE project aims at answering these questions by providing ways of evaluating and comparing risk assessment methods for security in ATM. The goal is to provide relevant stakeholders with the means to select the risk assessment methods that are best suited for the task at hand, for example, security assessment in relation to introduction of a particular new system by taking into account a specific aspect of security.

This paper outlines the initial version of the EMFASE empirical framework. The framework uses the Method Evaluation Model (MEM) proposed by Moody [1] for investigating the value of a security risk assessment method.

According to Moody, methods have no 'implicit' value, only pragmatic: a method in general and a risk assessment method in particular, cannot be true or false, but rather effective or ineffective. The objective of the EMFASE framework, therefore, is not to demonstrate that a 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" [1]. 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 enhancing the quality of the result.

The paper is organized as follows. Section II introduces a set of success criteria for security risk assessment methods in the ATM domain, and relates the criteria to the constructs of the MEM. The MEM constructs incorporate the aspects of a successful method, and the EMFASE framework helps to identify which of the success criteria contributes to which MEM constructs, and why. The first EMFASE empirical framework is presented in Section III, while Section IV provides an overview of the results obtained from empirical studies conducted so far. Section V concludes the paper and gives some insights on future work.

II. SUCCESS CRITERIA FOR ATM SECURITY RISK ASSESSMENT METHODS

In this section we introduce the success criteria that we have identified in order to evaluate and compare methods for security risk assessment within EMFASE.

A. *Success Criteria Identification Process*

We carried out an initial survey among ATM stakeholders to identify success criteria for security risk assessment (SRA) methods during the 6th Jamboree of the SESAR project 16.06.02 (Security support and coordination function), held in Brussels on 12 November 2013. The raw data collected during the survey were analysed through coding techniques drawn from grounded theory [2]. After this analysis a first set of high-level success criteria was identified. The criteria were reviewed, categorized and complemented by security experts in the EMFASE consortium. Subsequently, we further analysed the identified success criteria in order to relate them to the Method Evaluation Model (MEM).

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 success criteria and their relationship with MEM constructs will be further investigated and validated during the course of the EMFASE project. **Figure 1** summarizes the process carried out to identify the success criteria during the first project year. In the continuation of the project we will revise the identified criteria, their categorization, and their relations to the MEM constructs based on new insights and the knowledge gathered during EMFASE experiments.

B. *Coding of Survey Results*

The survey included a questionnaire that was filled in by the participants individually, as well as focus group interviews with the participants. 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 required to consider security risks 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.

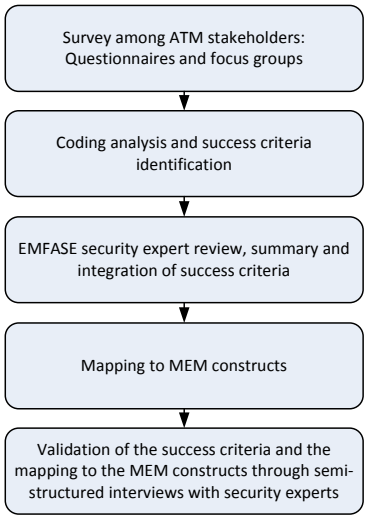

Figure 1: EMFASE criteria identification process

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* [2] which is a content analysis technique. We first 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. The identified codes were grouped by their similarity and classified into categories. For each category we counted the number of statements as a measure of their relative importance. We employed multiple coders working independently on the same data and then compared the results. This was to minimize the chance of errors from coding and increases the reliability of results.

C. *Identified Success Criteria*

TABLE 1 summarizes the main criteria reported by the professionals. We considered as the main identified criteria only the ones for which the participants made at least ten statements. We can observe here that while the main bulk of the statements fall into six 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.

TABLE 1: OCCURRENCES OF REPORTED SUCCESS CRITERIA

CRITERION	N° OF STATEMENTS
Clear steps in the process	28
Specific controls	24
Easy to use	19
Coverage of results	14
Tool support	13
Comparability of results	10
Others	
- 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**

Guided by the identified criteria, EMFASE security experts identified further method features or artefacts that could contribute to fulfil the criteria. 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: Compliance with ISO/IEC standards, a well-defined terminology, documentation templates, modelling support, visualization, systematic listing

of results, practical guidelines, assessment techniques, lists and repositories, and 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, as shown in **TABLE 2** are used for structuring the EMFASE empirical framework.

TABLE 2: SUCCESS CRITERIA AND CATEGORIES

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

D. *Success Criteria and Risk Assessment Methods Evaluation Model*

The Method Evaluation Model (MEM) considers three dimensions of "success",: *actual efficacy*, *perceived efficacy* and *adoption in practice* as shown in **Figure 2**.

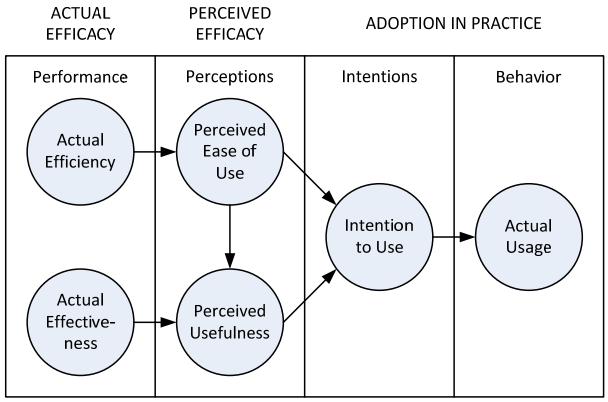

Figure 2: Method Evaluation Model

Actual efficacy is given by **actual efficiency,** which is the effort required to apply a method, and by **actual effectiveness** which is the degree to which a method achieves its objectives. Perceived efficacy is broken down in **perceived ease of use**, which is the degree to which a person believes that using a particular method would be free of effort, and **perceived usefulness,** which is the degree to which a person believes that a particular method will be effective in achieving its intended objectives. Adoption in practice is expressed by **intention to use** that is the extent to which a person intends to use a particular method, and **actual usage**, that is the extent to which a method is used in practice. The arrows between the constructs in **Figure 2** depict the expected *causal relationships* between the constructs. For example, one can expect that perceived usefulness is determined by actual effectiveness and perceived ease of use.

In **TABLE 3** 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. For example, the use of catalogues of threats and security controls is expected to improve the perceived ease of use, the perceived usefulness and the actual efficacy (efficiency and effectiveness).

The relations between the success criteria and the success constructs as presented in **TABLE 3** are our initial hypotheses about which criteria contributes to which MEM success constructs. The EMFASE framework and experiments investigate our initial hypotheses, and develop causal explanations based on the results of the empirical studies.

TABLE 3: SUPPORTING CRITERIA AND PARAMETERS IN RELATION TO THE MEM SUCCESS CONSTRUCTS

	SUCCESS CONSTRUCTS (MEM)			
SUPPORTING CRITERIA	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

III. THE EMFASE FRAMEWORK

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

A. *Purpose and Target Group*

The intended target group of the EMFASE framework is personnel that are responsible for developing the security cases for 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 and other ATM security personnel in addressing ATM security, and in conducting the security activities as specified by SESAR ATM Security Reference Material provided by project 16.06.02 [3]. The security activities include conducting security risk assessments and identifying adequate security controls for unacceptable risks. Moreover, for the ATM security personnel to effectively and efficiently conduct the security risk assessments the mentioned security reference material, as well as the European Operational Concept Validation Methodology (E-OCVM) [4], should include guidance on which SRA methods to use. EMFASE aims 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.

B. *Empirical framework*

In the following we present the EMFASE empirical framework, which includes a framework scheme and a protocol for conducting the experiments.

1) *Framework scheme*

The scheme for the initial EMFASE empirical framework is shown in **TABLE 4**. In the following we explain its content step by step.

The first column (#) refers to the EMFASE experiments that we have conducted or that are to come. 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 under four variables: 1) **Method experience** indicates whether (Y) or not (N) the participants of the experiment have prior experience with the SRA methods object of study; 2) **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; 3) **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; 4) **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 and to the identified SRA method success criteria as categorized in **TABLE 2**. 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, studies can be conducted in which the time to complete a task and produce a result is fixed and limited. 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 remaining columns refer to the SRA success criteria presented in **TABLE 2**.

TABLE 4: FRAMEWORK SCHEME

#	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		

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 rows in **TABLE 4** 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.

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 to the framework scheme and the success criteria. This protocol was used in conducting the EMFASE experiments 1 to 4 as listed in **TABLE 4**. Conceptually, the protocol is divided in two parallel streams that are merged in time as shown in Figure 3, namely an execution stream (E) and a measurement stream (M).

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

The **measurement stream** gathers the quantitative and qualitative data that will be used to evaluate the methods. Similarly to the execution stream, it consists of three phases: a) **Training**: Participants are administered a demographic questionnaire (**M1**). Participants are then distributed a post-training questionnaire to determine their initial perception of the methods and the quality of the tutorials (**M2**). M1 targets the threat to internal validity represented by participants' previous knowledge of the other methods; b) **Application**: The participants are requested to answer a post-task questionnaire about their perception of the method after each application session (**M3**). c) **Evaluation**. The participants' perception and feedback on the methods are collected through post-it note sessions, and focus group interviews (**M4**). The participants are also requested to answer a post-task questionnaire about the quality of the organization of the empirical study (**M5**). Furthermore, the method designers evaluate whether the groups of participants have applied the method correctly (**M6**), while domain experts assess the quality of identified threats and security controls (**M7**). The last two steps address two issues that may affect both conclusion and construct validity. Indeed, any method can be effective if it does not need to deliver useful results for a third party.

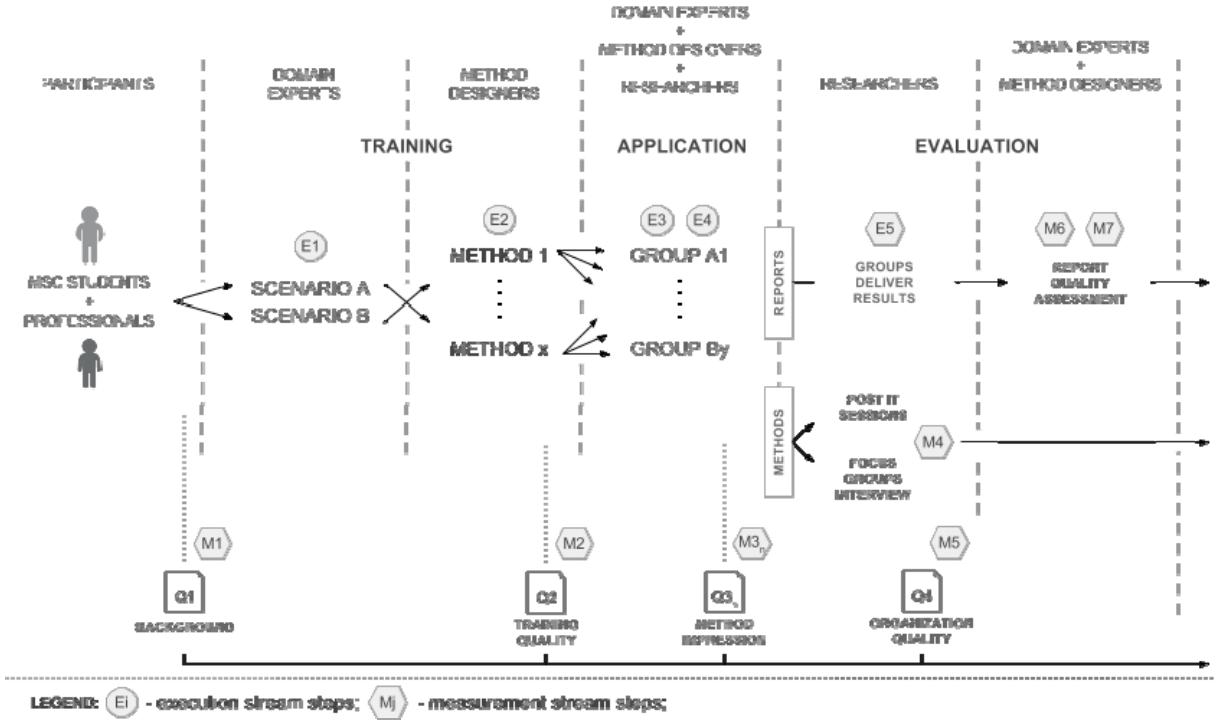

Figure 3: Empirical protocol to compare two SRA methods

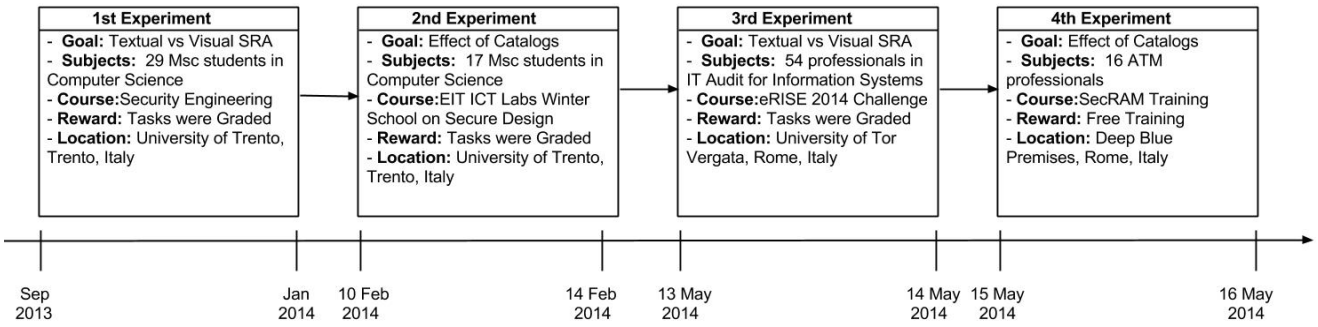

Figure 4: Empirical studies timeline

I. 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 previous section. As shown in Figure 4, 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 of these types of studies was first conducted with MSc students (Experiment 1 and 2) and then with Professionals (Experiment 3 and 4). The results of the 3rd and 4th experiment with security and ATM professionals are still under analysis.

A. *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 [6] as instance of a textual method.

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: 1) **Training.** Participants were given a 2 hours tutorial on the Smart Grid application scenario (not ATM-related to better test SRA generality and customizability) and a 2 hours tutorial on the visual and textual methods. Subsequently the participants were administered a questionnaire to collect information about their background and their previous knowledge of other methods. 2) **Application.** Once trained on the Smart Grid scenario and the methods, the participants had to repeat the application of the methods on two different aspects, namely Network and Database Security and Web Application Security. They were allowed to deliver intermediate presentations and reports to get further feedback. At the end of the course, each participant submitted a final report documenting the application of the methods on the two aspects. 3) **Evaluation**. 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. In addition, after the 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

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 (top) shows that the textual method did better than the visual one in identifying threats. But the results of the Friedman test do not show any significant differences in the number of threats among neither all threats (Friedman test returned *p-value* = 0.57) nor specific threats (Skillings–Mack test returned *p-value* = 0.17). In contrast, Figure 5 (bottom) shows that the visual and textual methods produced the same number of security controls. This is attested also by the results of statistical tests, which show that there is no statistically significant difference in the number of security controls of neither all security controls (Friedman test returned *p-value* = 0.57) nor the specific security controls (ANOVA test returned *p-value* = 0.72). Thus, we can conclude that there was no difference in the actual effectiveness of the visual and textual method for security risk assessment in this particular experiment.

Participants' Perception. The average of responses shows that the participants preferred the visual method over the textual method with statistical significance (Mann-Whitney test returns Z = -5.24, p-value = $1.4 * 10^{-7}$, es = 0.21).

Perceived Ease of Use. The visual method did 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.

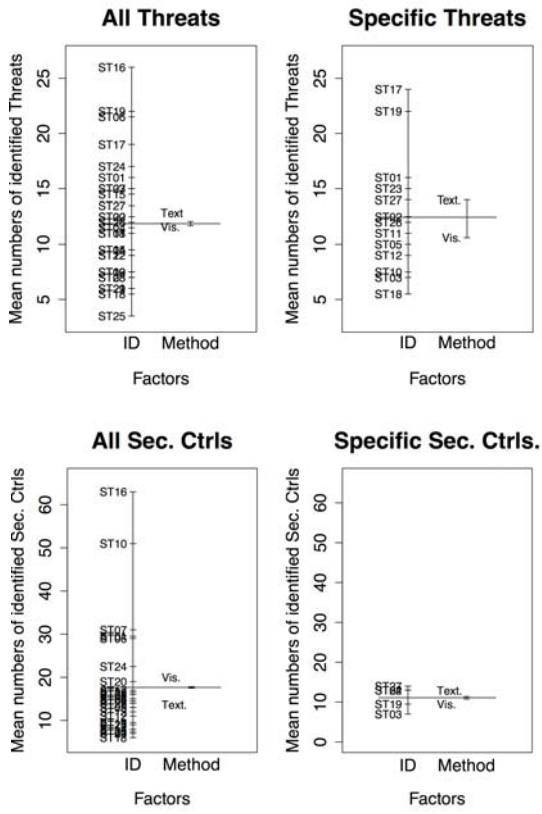

Figure 5: Actual Effectiveness: Number of threats and security controls

Perceived Usefulness. The visual method did 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 did 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).

We can conclude that in this experiment, overall, the visual method was preferred over the textual one with statistical significance. The difference in the perception of the visual and textual methods can likely be explained by the differences between the two methods; the diagrams used by the visual method help participants in identifying threats and security controls because they give an overview of the threats that harm an asset, while using tables makes it difficult to keep the link between assets and threats.

B. *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 [7]. 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 [8].

1) *Experimental procedure*

The experiment was held in February 2014 and organized in three main phases: 1) **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 one of two sets of catalogues of threats and security controls. The participants were given a tutorial on the method application of the duration of 8 hours spanned over 2 days. The tutorial was divided into different parts. Each part consisted of 45 minutes of training of a couple of steps of the method, followed by 45 minutes of application of the steps and 15 minutes of presentation and discussion of the results with the expert. 2) **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. 3) **Evaluation**. The 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.

2) *Experimental results*

To avoid bias in the evaluation of SESAR SecRAM and of the catalogues, we asked three security experts in security of the ATM domain to assess the quality of the 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.

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 groups and p-value = 0.94 (t(6) = −0.08, Cohen's d = 0.06) for good groups.

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 and good groups. However, Mann-Whitney test shows that this difference is not statistically significant in case of all groups ($Z = -0.74$, p-value = 0.56, $r = -0.24$) and good groups ($Z = -1.15$, p-value = 0.34, $r = -0.41$).

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

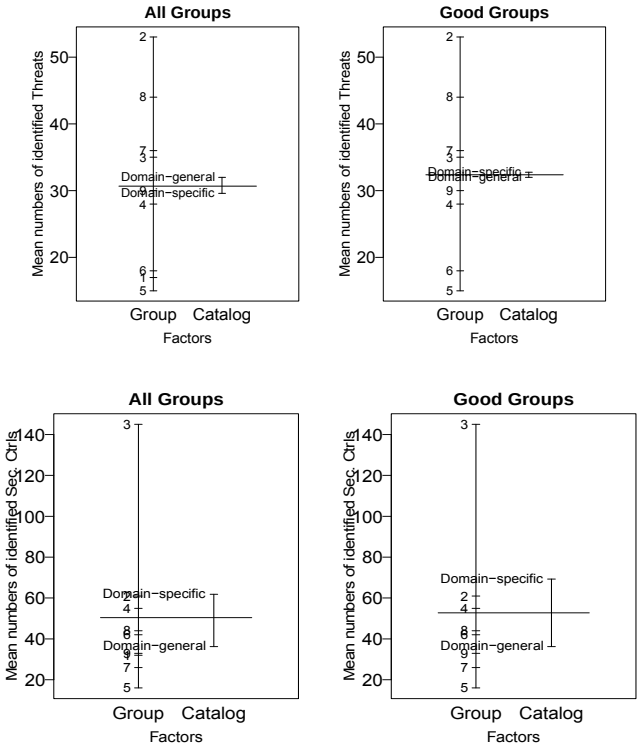

Figure 6: Actual effectiveness

Method's Perception. The overall perception of the method is higher for the participants that applied domain-specific catalogues with statistical significance for both all (Mann-Whitney (MW) test returned: $Z = -3.97$, $p = 7 * 10-5$, es = 0.17) and good participants (MW returned: $Z = -2.31$, $p = 0.02$, es = 0.10). The same results hold for perceived usefulness of the method: we have a statistically significant difference (MW returned: $Z = -2.57$, p-value = $7.3 * 10^{-3}$, es = 0.61) for all participants and good participants (MW returned: $Z = -2.31$, p-value = 0.02, es = 0.10). For perceived ease of use and intention to use the MW test did not reveal any statistically significant difference both for all and good participants.

In summary, 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. However, the overall perception and perceived usefulness of the method is higher when used with the domain-specific catalogues, which are considered easier to use than the domain-general ones.

II. CONCLUSIONS

In this document we have presented the first version of the EMFASE empirical evaluation framework and summarized the results obtained from the empirical studies conducted so far in. The studies indicate that visual methods for security risk assessment are better perceived than textual ones, and that the perceived usefulness of security risk assessment methods is higher when used with domain-specific catalogues

The EMFASE consortium is designing and organizing new studies to enrich and complement the ones already carried out, and to further validate the framework itself. We will conduct an experiment where we investigate comprehensibility of graphical versus tabular notations to represent risk models. We are also designing and preparing direct observations of professionals applying SRA methods in their daily work.

ACKNOWLEDGMENT

This work is co-financed by EUROCONTROL acting on behalf of the SESAR Joint Undertaking (the SJU) and the EUROPEAN UNION as part of Work Package E in the SESAR Programme. Opinions expressed in this work reflect the authors' views only and EUROCONTROL and/or the SJU shall not be considered liable for them or for any use that may be made of the information contained herein. This work has been partly supported by the EU under grant agreement n.285223 (SECONOMICS).

REFERENCES

[1] 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

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

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

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

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

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

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

[8] SESAR 06.09.03: OSED for Remote Provision of ATS to Aerodromes, Including Functional Specification. Deliverable D04, 2014

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

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