Which Security Catalogue Is Better for Novices?

Katsiaryna Labunets, Federica Paci, Fabio Massacci
DISI, University of Trento, Italy
Email: {name.lastname}@unitn.it

Abstract—**Several catalogues of security threats and controls have been proposed to help organizations in identifying critical risks and improve their risk posture against real world threats. But the role that these catalogues play in a security risk assessment has not yet been investigated. In this paper we report an experiment with 18 MSc students conducted to compare the effect of using domain-specific and domain-general catalogues of threats and security controls on the actual efficacy and perception of a security risk assessment method. The experimental results show that there is no difference in the actual efficacy of the method when applied with the two types of catalogues. In contrast, the perceived usefulness of the method is higher for the participants who have used the domain-specific catalogues. In addition, the domain-specific catalogues are perceived as easier to use by the participants.**

Index Terms—**empirical study, controlled experiment, security risk assessment methods, threats catalogues, security controls catalogues, Method Evaluation Model (MEM)**

I. Introduction

In the last decade several methods, frameworks and standards to identify and analyze threats and security controls in the early phases of the system development life-cycle have been proposed – ISO 27005 [1], NIST 800-30 [2], STRIDE [3], SABSA [4]. All these methods provide catalogues of threats and security controls to help organizations in identifying critical risks and improve their risk posture against real world threats. Indeed, using catalogues should facilitate the identification of threats and security controls during a security risk assessment. To test this hypothesis we conducted an initial study [5] with Air Traffic Management (ATM) professionals. Surprisingly, the study showed that non-security experts conducting a security risk assessment with catalogues identified threats and security controls of similar quality of the one identified by security experts without catalogues. Therefore, we conducted a second study to further investigate how non-security experts (e.g. MSc students) perform when using catalogues. The study aims to investigating the effect of using catalogues on the actual efficacy and perception of a security risk assessment method. In particular, we assessed the effect of using domain-specific catalogues versus domain-general ones. The dependent variables were the *actual efficacy* of the method measured as number and quality of threats and security controls and the participants' *perceived ease of use* and *perceived usefulness* of the method and the two type of catalogues. The independent variables were the *method* and the *catalogues*.

The experiment involved 18 MSc student from different majors in Computer Science. They were divided into 9 groups half of which applied a security risk assessment method with the domain-specific catalogues and the other half with the domain-general catalogues. Each group analyzed an emerging operational concept in the ATM domain called Remotely Operated Tower (ROT). The method selected was ATM Security Risk Assessment Method (SecRAM), a security risk assessment method used in the ATM domain. The method has been designed for users who have no expertise in security and thus it provides catalogues of threats and security controls specific for the ATM domain. Instead, as domain-general catalogues we used the catalogues of threats and controls of the BSI IT-Grundschutz standard.

The results indicate that both types of catalogues have no significant effect on the actual efficacy of the method. In particular, there is no statistically significant difference in the number and quality of threats and security controls identified with the two types of catalogues. However, the perceived usefulness of the method is higher when used with the domain-specific catalogues. In addition, the domain-specific catalogues are perceived as easier to use than the domain-general ones.

The rest of the paper is structured as follows. In the next section we present the related work (§II). Then, we describe our research approach (§III) and present the settings of the study (§IV). Further, we present the results (§V) and summarize the main findings (§VI). Finally, we discuss threats to validity (§VII) and conclude the paper (§VIII).

II. Related Work

In this section, we first discuss empirical studies on the evaluation of security risk assessment methods and then we report empirical studies on the reuse of security knowledge.

Evaluation of Security Risk Assessment Methods: In the realm of methods for eliciting threats and security controls, there are only few papers [6], [7], [8], [9], [10], [11], [12] that evaluated whether these methods work in practice. Most of them based the evaluation on the Method Evaluation Model (MEM) [13] which provides constructs to measure methods success: *actual efficiency*, *actual effectiveness*, *perceived ease of use (PEOU)*, *perceived usefulness (PU)*, and *intention to use (ITU)*. For example, Opdhal and Sindre [6] carried out two controlled experiments (28 and 35 students) to compare two methods for threats identification, namely attack trees and misuse cases. In [11] Opdhal and colleagues repeated the experiment with industrial practitioners. Both experiments showed that attack trees help to identify more threats than misuse cases. Similar controlled experiments with students were reported by Stålhane et al. in [14], [9], [10], [8] where misuse cases were compared with other approaches for safety and security. Stålhane et al. [14] reported an experiment with 42 students where they compared misuse cases to Failure Mode and Effects Analysis (FMEA) in analyzing use cases. They found that misuse cases are better than FMEA for analyzing failure modes related to user interactions. In a similar setting [9], the authors compared misuse cases based on use case diagrams to those based on textual use cases. The results of the experiment with 52 students showed that textual use cases produce better results due to more detailed information.

The e-RISE challenge organized by the University of Trento [7] reported an interesting protocol to perform empirical comparisons of different risk assessment methods by using both practitioners and students. The challenge revealed that threat-based methods perform better for security analysis. More recently, Labunets et al. [12] adopted a similar experimental protocol to conduct a controlled experiment with 28 MSc students to compare two types of security risk assessment methods, visual (CORAS) and textual (SREP) methods. The results showed that visual methods are more effective in identifying threats and better perceived than the textual ones. In the current experiment we adopted an experimental procedure similar to the one proposed by Labunets et al. In addition, we limited threats to conclusion validity because *a)* participants were trained by a EUROCONTROL expert who usually trains professionals working in the ATM domain and *b)* the participants had two full days to apply the method to a new ATM operational concept. Thus, our experiment is high on realism.

Most of these experiments have some limitations. Experiments such as [6], [8], [9], [10] involve students and usually have a short duration (less than two hours). This may introduce threats to conclusion and external validity. Conclusion validity can be biased because subjects do not have enough time to understand the application scenario and to fully apply the methods under evaluation. Further, if the time for the execution of the experiment is short, it is impossible to use a realistically-sized application scenario. Hence, the experiments lack realism. The experiments in [7], [11] mitigate threats related to experiment duration, scenario's complexity and participants' experience because they last several days and include practitioners. The experiment by Labunets et al. counterbalances the use of students as participants with the duration of the experiment that lasted several weeks rather than just two hours and the use of a real application scenario.

Effect of Reusing Security Knowledge: To the best of our knowledge there are few papers [15], [5] that aim to investigate the effect of reusing security knowledge. Yskout et al. [15] investigated the effect of using the catalogue of security patterns on the quality of a security design and productivity of the designers. The study involved 64 MSc students who worked in teams of 2 members. The study adopted a within-subject design such the teams conducted security analysis both with and without security pattern catalogue. The results showed that there was no difference in the quality of results and productivity between teams who used security patters and who does not. However, participants preferred to do security analysis with the support of security pattern. Similar to this study we also investigate the effect of using catalogues of threats and security controls on the quality of results. However, we assessed quality differently. Yskout et al. measured quality of a security design as a number of covered misuse cases per task. While we rely on the quality assessment by independent security experts.

In a previous study with professionals in the ATM domain [5], we investigated the effect of using catalogues on the *actual effectiveness* and *perception* of a security risk assessment run by non-expert (with the catalogues) and by expert (without a catalogues). Actual effectiveness was quantitatively investigated as the quality of threats and security controls identified by the participants. Perception was assessed both quantitatively via post-task questionnaire and qualitatively via focus group interviews with the participants.

The main findings are that professionals that are not security experts can obtain almost the same results of domain experts without a catalogue while applying a security risk assessment methods. A domain-specific catalogue was perceived to be slightly more useful than a domain independent one, in particular because the former allowed a better navigation to non-expert users. Based on this results, we decided to conduct a second experiment with non-experts in order to investigate better the effect of domain-specific and domain-general catalogues on security risk assement's efficacy and perception.

III. Research Method

The goal of the experiment is to compare the effect of using *domain-general* versus *domain-specific* catalogues of threats and security controls on the actual efficacy and perception of a security risk assessment method. Table I reports the list of hypothesis to be tested. The dependent variables are *actual efficacy* of the method, *perception of the method* and *perception of the catalogues*. Perception is broken down in *PEOU* and *PU*.

We measured actual efficacy as the number and quality of threats and security controls produced by the participants. Two researchers independently extracted and counted the number of threats and security controls included in groups of participants' final reports. We asked three experts in security of ATM domain to assess the quality of threats and security controls identified by the participants. The experts 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 threats/security controls are not specific for the scenario; *Fair* (3), when *some* of them are related to the scenario; *Good* (4), threats/security controls are specific for the scenario; and *Excellent* (5), when the threats are significant for the scenario and security controls propose real solution for the scenario. Based on this scale the groups who achieved an assessment higher than *Fair* were classified as *good groups*.

TABLE I: List of hypothesis

ID	Hypotheses
$H1_0$	No Difference in the number of threats found with domain-specific and with domain-general catalogue
$H2_0$	No Difference in the number of security controls found with domain-specific and with domain-general catalogue
$H3_0$	No Difference in the quality of threats found with domain-specific and with domain-general catalogue
$H4_0$	No Difference in the quality of security controls found with domain-specific and with domain-general catalogue
$H5_0$	No Difference in the participants' PEOU of method when used with domain-specific and with domain-general catalogue
$H6_0$	No Difference in participants' PU of method when used with domain-specific and with domain-general catalogue
$H7_0$	No Difference in the participants' PEOU of using domain-specific and domain-general catalogue
$H8_0$	No Difference in participants' PU of using domain-specific and domain-general catalogue

We measured PEOU and PU of the method and of the catalogues by means of a post-task questionnaire inspired to the Method Evaluation Model (MEM) [13]. The questions were formulated in opposite statements format with answers on a 5-point Likert scale. The questionnaire also included open question to give participants the opportunity to provide feedback on the method and the catalogues. The post-task questionnaires are reported in [16].

A. Context Selection

Below we present the application scenario analyzed by the participants and the method and catalogues applied to identify threats and security controls for the scenario.

Method and Catalogues Selection: We selected the SESAR ATM Security Risk Assessment Method (SecRAM) [17] as security risk assessment method to be applied by the participants for three main reasons: *a)* it is a method used in the ATM domain to conduct security risk assessment of operation concepts; *b)* the application of SecRAM is supported by the use of catalogues and threats and controls; and *c)* a SecRAM expert was available to train our participants. SecRAM is developed within the SESAR JU project 16.02.03 (Security Risk Assessment – Security Risk Assessment Methodology) with the goal of providing a method that is applicable to all ATM Operational Focus Areas (OFAs), that is understandable to personnel with little expertise and background in security and risk management, and that allows security risk assessment results from different OFAs to be compared. The SecRAM process is divided into seven steps as follows: 1) primary asset identification and impact assessment, 2) supporting assets identification and evaluation, 3) threats scenarios identification, 4) impact evaluation, 5) likelihood evaluation, 6) risk level evaluation, and 7) risk treatment. As shown in Figure 1 tables are used to represent the results of each step's execution.

Since SecRAM comes with catalogues of threats and security controls to support non expert personnel, we used them as an instance of *domain-specific catalogues* (DOM CAT). These catalogues were developed by EUROCONTROL to provide the best practices in security and safety analysis for ATM domain. They consist of three main parts: threats, pre and post security controls. The catalogues include 32 generic threats of three types: Physical, Information and Procedural. For each generic threat there is at least an example of corresponding specific threat, its potential impact and evidence of the threat. The catalogues also propose a number of pre and post controls to mitigate each threat. They contain 33 pre and 18 post countermeasures for the threat part. Each control contains the link to the mitigated threats and a description of the procedure: in case of pre control the catalogues provide a description of how to prevent the threat, in case of post controls they specify a response to the threat after-effects and recovery plan.

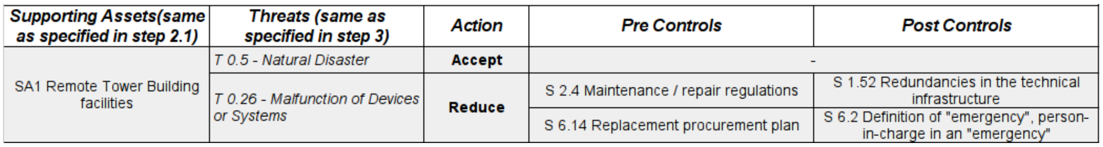
Instead, we chose as an instance of *domain-general catalogues* (GEN CAT) the threats and security controls catalogues of the BSI IT-Grundschutz standard [18]. This standard is developed by Bundesamt fr Sicherheit in der Informationstechnik (BSI – Federal Office for Information Security (English)), and it is widely used in Germany. It is compatible with the ISO 2700x family of standards. The BSI IT-Grundschutz catalogues not only describe possible threats and what has to be done in general to mitigate them, but they also provide concrete examples on how security controls should be implemented. The catalogues describes threats of the following types: Basic threats (46 threats), Force Majeure (19 threats), Organizational Shortcomings (174 threats), Human Error (116 threats), Technical Failure (89 threats) and Deliberate Acts (177 threats). The safeguards catalogues describes the following types of countermeasures: Infrastructure (80 controls), Organization (515 controls), Personnel (90 controls), Hardware and software (435 controls), Communication (173 controls) and Contingency planning (151 controls). Hence, it is clear that BSI IT-Grundschutz catalogues cover very wide spectrum of security and safety problem.

Application Scenario: As application scenario to be used by the participants, we chose a new operational concept which is emerging in the ATM named Remotely Operated Tower (ROT). ROT is a technical solution deployed at small and medium-sized airports, which enables an airport tower to be remotely operated via a digital network without human controllers on-site. A set of 360 cameras, sensors and surveillance radars located at the aerodrome provides a 360-degree real-time view of the airports and exhaustive information. This data is used by Air Traffic Control and/or Aerodrome Flight Information Services Operators at ROT centers which remotely control different airports at one time.

IV. Experimental Design and Execution

In this section we discuss the experimental design, protocol and execution of the experiment.

Experimental Design: We chose a *between-subject design* where participants work in group of two and apply the security risk assessment method with one of two types of catalogues. Nine groups were randomly assigned to the catalogues: four groups applied SESAR SecRAM method to the ROT scenario using GEN CAT catalogues while the other five groups used DOM CAT catalogues.



Supporting Assets(same as specified in step 2.1)	*Threats (same as specified in step 3)*	*Action*	*Pre Controls*	*Post Controls*
SA1 Remote Tower Building facilities	*T 0.5 - Natural Disaster*	**Accept**	-	
	T 0.26 - Malfunction of Devices or Systems	**Reduce**	S 2.4 Maintenance / repair regulations	S 1.52 Redundancies in the technical infrastructure
			S 6.14 Replacement procurement plan	S 6.2 Definition of "emergency", person-in-charge in an "emergency"

Fig. 1: SECRAM - Selection Controls Table

TABLE II: Participants' Demographic Statistics

Variable	**Scale**	**Mean**	**Distribution**
Age	Years	25.06	33% were 21-24 years old; 67% were 25-29 years old
Gender	Sex		56% male; 44% female
Education Length	Years	5.17	44% had <5 years; 6% had 5 years; 50% had >5 years
Work Experience	Years	2.90	33% had no experience; 22% had 1-2 years; 44% had 3-5 years
Experience in Security/Privacy Initiatives	Yes/No	-	28% involved; 72% not involved
Level of Expertise in Safety Technology	1(Novice)-5(Expert)	1.83	44% novices; 28% beginners; 28% competent users
Level of Expertise in Safety Regulation and Standards	—"—	1.56	61% novices; 22% beginners; 17% competent users
Level of Expertise in Security Technology	—"—	2.28	17% novices; 50% beginners; 22% competent users; 11% proficient users
Level of Expertise in Security Regulation and Standards	—"—	1.89	33% novices; 44% beginners; 22% competent users
Level of Expertise in ATM	—"—	1.06	94% novices; 6% beginners

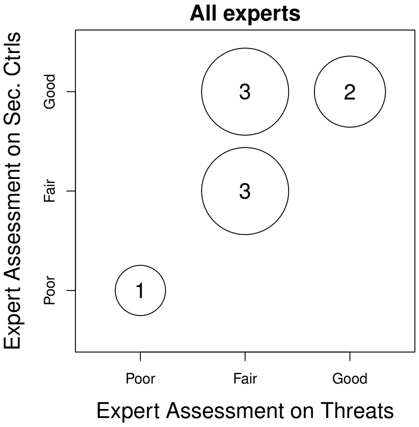


Fig. 2: Overall Expert Assessment of Quality of Threats and Security Controls for Groups of Participants

Experimental Procedure: The study was based on the step-wise process consisting of 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 Remotely Operated Tower of the duration of 1 hour. After the tutorial, participants were divided into groups and received the training material. The training material consists of a detailed description of the scenario and the two catalogues. Since the DOM CAT catalogues are confidential material for EUROCONTROL, the participants received only a paper version of the catalogues and had to sign a non-disclosure agreement.

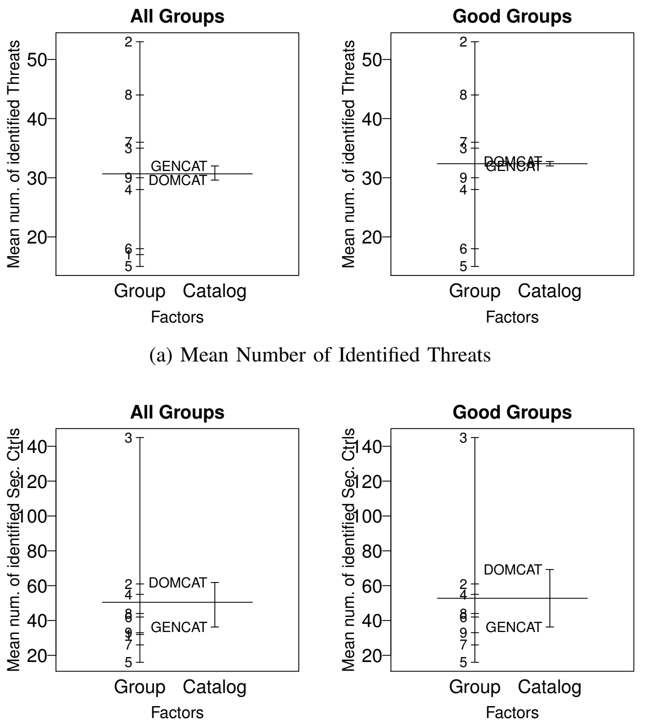
Then, the participants were given a tutorial on SESAR SecRAM method of the duration of 8 hours spanned over 2 days. The tutorial was divided into different parts. Each part consisted of 45 minutes of introduction of a couple of steps of the method, followed by 45 minutes of application of the steps and 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 revise the security risk assessment. After the application phase participants delivered their final reports documenting the conducted security risk assessment of the ROT. Then a post-task questionnaire was administered to the participants to collect their perception of the method and the catalogues.

Evaluation. Three experts independently judge the quality of the threats and security controls identified by the groups of participants, providing marks and comments.

The post-task questionnaires were inspired by the Technology Acceptance Model (TAM) [19]. The questions were formulated in opposite statements format with answers on a 5-point Likert scale (1 - Strongly agree with left statement; 2 - Agree with left statement; 3 - Not certain; 4 - Agree with right statement; 5 - Strongly agree with right statement). To prevent participants from "auto-pilot" answering, a half of the questions were given with the most positive response on the left and the most negative on the right. The post-task questionnaires are reported in [16].

Participants Demographics: The experiment was held in February 2014 at the University of Trento. The participants of the experiment were 18 MSc students from different universities in Europe participating to EIT ICT Labs, a partnership between universities, research center and companies that promotes innovation in education and research. Table II presents descriptive statistics about the participants. Most of the participants (44%) reported that they had at least 3 years of working experience, some participants (22%) reported ≤ 2 years of workings experience, and the rest did not report any working experience. Also some participants (28%) reported that they have been involved in security/privacy initiatives, the rest did not report any similar experience. With respect to the knowledge in safety technologies, safety and security regulation and standards, our participants had limited expertise, while in security technologies they reported an extensive general knowledge. Our participants also had no prior knowledge of the ATM domain.



(a) Mean Number of Identified Threats

(b) Mean Number of Identified Security Controls

Fig. 3: Mean Number of Identified Threats and Sec. Controls

V. Results

In this section we report the results on method's actual efficacy and perception of the method and the catalogues.

Actual Efficacy: As mentioned before the actual efficacy was measured as a number and quality of threats and security controls identified by the groups. Two researchers independently counted the number of threats and security controls identified by the groups. The quality of the threats and controls was evaluated independently by three ATM security experts. The three experts reported a similar evaluation for each group. Figure 2 illustrates the average of experts' evaluation for threats (reported on x-axis) and security controls (on y-axis). Only one group out of nine performed poorly. In what follows, we compare the results produced by all groups with the one of good groups but we draw our conclusions only upon the results of good participants.

We investigated whether there is a difference in the number and quality of threats and security controls identified with each type of catalogues. To analyze the difference in the number of threats we used unpaired t-test. Mann-Whitney (MW) test we used *a)* for the number of security controls because this sample failed equal variance assumption and *b)* for the quality of threats and security controls because these are ordinal data. To calculate the effect size for t-test we used [20], for MW test effect size we used formula r=$Z_{MW}/\sqrt{N}$, where N is total number of observations.

Figures 3a and 3b show the difference in the number of threats and security controls identified with two types of catalogues by all groups and good groups only. As wee can see, there is no difference in the number of threats between the groups applied the method with DOM CAT or GEN CAT catalogues (t-test results: *t=-0.26, p=0.8, Cohen's d=0.18*). This is also supported by the results of the good groups (*t=-0.08, p=0.94, Cohen's d=0.06*). In case of the number of security controls we can observe more evident difference in favor of DOM CAT catalogues. However, the results of MW test did not support it (MW test results: *Z=0.73, p=0.56, r=0.24* for all and *Z=1.16, p=0.34, r=0.4* for good groups).

TABLE III: Groups, Their Results and Quality Assessment

Group ID	Catalogue	Quantity		Quality					
				Exp1		Exp2		Exp3	
		T	SC	T	SC	T	SC	T	SC
G01	DOM CAT	17	32	3	3	2	2	2	2
G02	DOM CAT	53	61	4	3	3	3	3	3
G03	DOM CAT	35	145	4	4	4	4	4	4
G04	DOM CAT	28	55	4	3	3	4	3	3
G05	DOM CAT	15	16	3	3	3	3	5	5
G06	GEN CAT	18	42	3	4	3	3	4	4
G07	GEN CAT	36	26	3	4	3	4	3	3
G08	GEN CAT	44	44	2	3	4	4	3	3
G09	GEN CAT	30	33	3	4	3	3	3	4

Table presents the information about number of threats (T) and security controls (SC) identified by groups and the assessment from three ATM experts on the quality of threats and security controls.

We also compared the quality of threats and security controls identified with the two types of catalogues. Table III reports the detailed results of risk assessment delivered by the participants and quality evaluations from three experts. As we can see, the quality of threats identified with DOM CAT catalogues (median threats quality is 3.33) is higher than the one of threats identified with GEN CAT catalogues (median is 3). In contrast, the quality of security controls identified with the support of DOM CAT catalogues (median is 3.33) is lower than the one of controls identified with GEN CAT catalogues (median is 3.67). However, the results of MW test on overall quality across three experts show that these results are not statistically significant. MW test returned *Z=-0.74, p=0.24, r=0.42* for the overall quality of threats and *Z=0.77, p=0.52, r=0.26* for the overall quality of security controls.

Method's and Catalogues' Perception: The post-task questionnaire was analyzed to identify the difference in participants perception of the method applied with each type of catalogues and of the two type of catalogues. For the analysis of the method's and catalogue's perception we used the results collected from the good group participants. We believe that participants that were able to benefit from using catalogues could judge the usefulness of the method and catalogues. While the participants who identified threats and security controls of low quality are not the best candidates to reason about usefulness of the method and catalogue.

To analyze the results of post-task questionnaire we used non-parametric MW test due to ordinal type of the data. Before conducting analysis all responses were reverted to 5 being the best. Further we discuss the results for PEOU and PU questions about the method applied with catalogues and PEOU and PU questions asked directly about the catalogues. We also measured participants' ITU of the method and catalogues but

TABLE IV: Participants Responses to PU and PEOU Questions about Method and Catalogue (Good Participants)

(a) Questions about Method

	1	2	3	4	5	Total	Median
PU							
DOM CAT	0	9	6	13	4	32	4
GEN CAT	0	5	11	11	5	32	3
Total	0	14	17	24	9	72	-
PEOU							
DOM CAT	7	9	18	30	8	72	4
GEN CAT	1	9	33	19	10	72	3
Total	8	18	51	49	18	144	-

(b) Questions about Catalogue

	1	2	3	4	5	Total	Median
PU							
DOM CAT	1	7	39	47	10	104	4
GEN CAT	1	19	50	30	4	104	3.5
Total	0	26	89	77	14	208	-
PEOU							
DOM CAT	6	2	6	10	8	32	4
GEN CAT	0	12	11	9	0	32	3
Total	6	14	17	19	8	72	-

These tables report the total number of responses by the participants to PU and PEOU questions about the method and the catalogue. The columns describe response options on a scale 1-5 with 5 being the best option (1 - Strongly disagree; 2 - Disagree; 3 - Not certain; 4 - Agree; 5 - Strongly agree) and median value of the responses. The rows describe treatment group types: participants who applied method with a domain-specific catalogues (DOM CAT) and with domain-general catalogues (GEN CAT).

the results are not statistically significant. Table IVa reports the total number of responses by the participants to all method's PU and PEOU questions. There were 9 questions about method's PEOU and 13 questions about method's PU in the questionnaire. Similarly table IVb reports the total number of participants' responses to all catalogue's PU and PEOU questions. There were 4 PU and 4 PEOU questions in the catalogue post-task questionnaire. The detailed statistics of participants' responses are reported in Tables V and VI.

Method with Catalogues: The method has higher PEOU when used with DOM CAT catalogues than when is applied with the GEN CAT catalogues across good participants but the difference is not statistically significant. Similar results we have across all participants. This result can be illustrated by questions Q18 and Q19 about ease of evaluating the appropriateness of identified threats and security controls to the context which have the most significant difference.

The method has higher PU when used with DOM CAT catalogues than with GEN CAT catalogues across all and good participants with statistical significance. For example, the individual PU questions Q28 and Q29 about ease of comparison of threats and security controls identified with the method to the other methods have the most apparent difference supporting the overall PU result. We also measured participants ITU of the method applied with the catalogues but the results are not statistically significant.

Catalogues: DOM CAT catalogues have higher PEOU than GEN CAT catalogues (3 vs. 4 as median values). The result has only 10% significance for good participants, but it is supported across all participants with statistical significance. The most prominent questions for this result are questions Q3 about ease of use of the catalogues, and Q4 about easiness of finding specific threats with the catalogue. DOM CAT catalogues have higher PU than GEN CAT catalogues across all and good participants but with no statistical significance.

VI. Discussion

Actual Efficacy: There is no difference in the number and quality of threats and controls of the security risk assessment method when used with the domain-specific and the domain-general catalogues. Therefore, we cannot reject the null hypotheses $H1_0 - H4_0$. Similar results we received in the study with ATM professionals [5]. In order to identify statistically significant effect of catalogue type on the quality of results we would need to run experiment with at least 38 groups for the quality of threats and 101 groups for the quality of security controls (we used [21] to calculate sample size).

Method's Perception: The difference in the PEOU of the method with domain-specific or domain-general catalogues is negligible (3 vs. 4 as median values). The effect size is $-0.02 \in CI[-0.18, 0.15]$ and therefore, we would need more than 2968 participants to achieve a 80% power. The null hypothesis $H5_0$ cannot be rejected. Similarly in our previous study with ATM professionals we found no difference between participants' PEOU of the method applied with domain-specific or domain-general catalogues. In contrast, the PU of the method are higher when used with domain-specific catalogue and the difference is significantly larger (3 vs. 4) with the effect size equal to $-0.25 \in CI[-0.37, -0.11]$. To achieve 80% power we would need only 10 participants while we had 18. We can thus reject the null hypothesis $H6_0$. This finding differs from the results of our previous study with ATM professionals where there was no difference in the PU of the method when used with domain-specific or domain-general catalogues. This can be explained by the fact that participants of the first experiment who applied method with domain-specific or domain-general catalogues were domain but not security experts. Thus, for them when method applied with domain-specific or domain-general catalogues was equally useful. While the participants of the current experiment do not have neither domain nor security expertise. Therefore, applying security risk assessment method with domain-specific catalogues is perceived more useful than applying this method with domain-general method. For example, the participants of the current study made the following statements: "Provides guidelines, list of threats and security controls specific to the domain of the analysis" (DOM CAT participant) and "better than having nothing" (GEN CAT participants).

Catalogues' Perception: For PU the distinction between two catalogues is so small (3.5 vs 4 as median value) that we would need more than 746 participants to obtain a 80% power. The effect size is $0.05 \in CI[-0.2, 0.3]$. It can go in either directions. For PEOU the difference is much larger (3 vs. 4) and the effect size is negative $-0.24 \in CI[-0.46, 0.02]$. The confidence interval is essentially on the negative side.

A stronger result at 5% instead of 6% would only need 35 participants instead of 18. The higher PEOU of the domain-specific catalogues can be also explained by the feedback from the participants: “easy to find threats and controls – easy to understand relation between different kinds of control mechanism – easy to match all the information provided” (DOM CAT participant) and “a lot of controls are almost the same so it is difficult to differentiate then and decide which one is the best” (GEN CAT participant). Therefore, we cannot reject the null hypotheses $H7_0$ and $H8_0$.

VII. Threats to Validity

The main threats to validity of our study are related to conclusion and external validity [22]. The main threat to conclusion validity is related to the *sample size* that must be big enough to come to correct conclusions. We discussed the adequacy of the sample size together with the results in the previous section. As we can see for some variables we have enough data points (e.g., the PU of the method) or almost enough (e.g., the PEOU of the catalogues and the quality of threats), while for the other the effect size is so small that we would need to significantly increase the number of participants (e.g., the PEOU of the method and PU of the catalogues).

The main threat to internal validity could be *the size of catalogues* because domain-specific catalogues (155 pages) are significantly shorter than the domain-general catalogues (˜2500 pages). To mitigate this risk we prepare a short version of domain-general catalogues (˜55 pages) that contained only the list of available threats and security controls. But the participants still had access to the full version of the domain-general catalogues.

Another threat to conclusion validity is related to the *quality of threats and security controls* identified by the groups. We limited this threat by requesting three external domain experts to evaluate whether the threats and security controls identified by the groups were specific for the application scenario. Instead, the main threat to external validity is related to the *use of the students instead of practitioners*. We mitigate this threat by using MSc students which were close to finalize their education and start working in industry. Also at least one of the participants in each group has finished a course on security engineering. This allowed us to have groups with same level of expertise in security. In addition, to guarantee the quality of training we invited professional instructor from consulting company to train the participants on application scenario and method. Another threats is the *realism of experimental settings*. Our experiment had the duration of three days rather than a couple of hours like most of the experiment. This duration allowed us to use a complex enough application scenario and thus to generalize our results to the real projects.

VIII. Conclusion

Catalogues of threats and security controls should facilitate the identification of threats and controls. However, a previous study [5] that we conducted with ATM professionals showed that non-security experts who conduct a security risk assessment with catalogues identified threats and security controls of similar quality of the one identified by security experts without catalogues. Therefore, to understand better how non-security experts perform with catalogues, we conducted a second study with MSc students reported in this paper. The study focused on the effect of using domain-specific and domain-general threats and security controls catalogues on a security risk assessment method’s efficacy and perception. The results showed that the groups who used domain-specific and domain-general catalogues identified threats and security controls of a similar quality. These results support finding of our experiment with ATM professionals.

Another finding is that participants perceived as useful the use of the method with domain-specific (by the participants of the groups who produced good risk assessments). In contrast, in the previous experiment with ATM professionals where there was no difference in the perceived usefulness when method used with domain-specific or domain-general catalogues. However, when asked directly whether catalogues were effective such difference in perception was relatively small. For perceived ease of use we have an opposite situation. The participants found that the method was equally easy to use when used with the two types of catalogues. But when we asked directly about catalogues ease of use, the domain-specific catalogues were preferred to the domain-general ones.

Additional work has to be done to investigate the differences in the performance and perception of the method applied with different types of catalogues. Further replications required to increase the statistical significance of the results. The other domains and domain-specific catalogues should be taken into account in order to increase the generalizability of the findings.

Acknowledgment

This work has been partly supported by the SESAR JU WPE under contract 12-120610-C12 (EMFASE).

References

[1] ISO, “ISO/IEC 27005, Information technology Security techniques - Information security risk management,” Tech. Rep., 2011.

[2] G. Stoneburner, A. Goguen, and A. Feringa, “Risk management guide for information technology systems,” *Nist special publication*, vol. 800, no. 30, pp. 800–30, 2002.

[3] S. Hernan, S. Lambert, T. Ostwald, and A. Shostack, “Threat modeling-uncover security design flaws using the stride approach,” *MSDN Magazine-Louisville*, pp. 68–75, 2006.

[4] J. Sherwood, A. Clark, and D. Lynas, *Enterprise security architecture: a business-driven approach*. Backbeat Books, 2005.

[5] M. de Gramatica, K. Labunets, F. Massacci, F. Paci, and A. Tedeschi, “The role of catalogues of threats and security controls in security risk assessment: An empirical study with atm professionals,” in *Proc. of REFSQ ’15*. Springer, 2015, pp. 98–114.

[6] A. L. Opdahl and G. Sindre, “Experimental comparison of attack trees and misuse cases for security threat identification,” *Inf. Soft. Technology*, vol. 51, no. 5, pp. 916–932, 2009.

[7] F. Massacci and F. Paci, “How to select a security requirements method? A comparative study with students and practitioners,” in *Proc. of NordSec ’12*. Springer, 2012, pp. 89–104.

[8] T. Stålhane and G. Sindre, “Identifying safety hazards: An experimental comparison of system diagrams and textual use cases,” in *Proc. BPMDS ’12*, vol. 113, 2012, pp. 378–392.

TABLE V: Statistics of the Results of the Post-task Questionnaire about the Method

Q	Type	All participants							Good participants						
		GEN CAT			DOM CAT				GEN CAT			DOM CAT			
		Median	Mean	sd	Median	Mean	sd	Z_{MW}	Median	Mean	sd	Median	Mean	sd	Z_{MW}
1	ITU	4	3.5	0.76	3	3	0.82	1.34	4	3.5	0.76	3	2.75	0.71	1.9 ●
2	ITU	2.5	2.62	0.74	3	3	0.82	-1	2.5	2.62	0.74	3	2.75	0.71	-0.4
3	PEOU	3.5	3.62	0.74	4	3.4	1.17	0.09	3.5	3.62	0.74	3.5	3.25	1.28	0.44
4	PEOU	3.5	3.5	0.53	4	3.7	1.16	-0.97	3.5	3.5	0.53	4	3.62	1.3	-0.68
5	PEOU	4.5	4.25	0.89	4	3.5	0.97	1.58	4.5	4.25	0.89	4	3.62	0.92	1.33
6	ITU	3.5	3.5	0.93	3	3.2	0.63	0.78	3.5	3.5	0.93	3	3	0.53	1.29
7	ITU	3	3	0.93	3.5	3.3	0.82	-0.99	3	3	0.93	3	3.12	0.83	-0.51
8	ITU	3	3.25	1.04	3	3.3	0.67	-0.24	3	3.25	1.04	3	3.25	0.71	-0.11
9	ITU	3	3.25	0.89	4	3.5	0.71	-0.96	3	3.25	0.89	4	3.62	0.52	-1.27
10	PU	2.5	2.62	1.06	3.5	3.2	1.48	-0.96	2.5	2.62	1.06	3.5	3.25	1.28	-1.08
11	PEOU	4	3.88	1.36	4	3.8	1.03	0.42	4	3.88	1.36	3.5	3.75	1.16	0.38
12	ITU	3	3.12	0.64	3	2.7	0.95	1.01	3	3.12	0.64	3	2.75	1.04	0.74
13	PEOU	3.5	3.38	1.06	3.5	3.1	1.37	0.32	3.5	3.38	1.06	3	2.88	1.46	0.7
14	PU	3	3	0.53	3	3.5	0.97	-1.17	3	3	0.53	3	3.25	0.89	-0.51
15	PU	3	3.25	0.46	4	3.8	1.03	-1.35	3	3.25	0.46	3.5	3.62	1.06	-0.82
16	PU	3	3.25	0.46	4	3.7	0.67	-1.52	3	3.25	0.46	3.5	3.62	0.74	-1.11
17	PU	4	3.62	0.92	4	3.9	0.88	-0.52	4	3.62	0.92	3.5	3.62	0.74	0.17
18	PEOU	3	2.75	0.46	4	3.7	0.82	-2.57 **	3	2.75	0.46	4	3.5	0.76	-2.18 *
19	PEOU	3	3	0.53	4	3.7	0.82	-2.07 *	3	3	0.53	4	3.5	0.76	-1.62
20	PU	3	3.25	0.71	4	3.8	0.63	-1.57	3	3.25	0.71	4	3.62	0.52	-1.12
21	PU	4	3.62	0.52	4	3.8	0.92	-0.64	4	3.62	0.52	4	3.62	0.92	-0.12
22	PEOU	2.5	2.62	0.74	4	3.2	1.14	-1.36	2.5	2.62	0.74	3.5	3	1.2	-0.83
23	PEOU	3	3.5	0.76	3.5	3	1.25	0.47	3	3.5	0.76	3	2.75	1.28	1
24	ITU	3.5	3.38	1.06	3.5	3.7	0.82	-0.61	3.5	3.38	1.06	3	3.5	0.76	-0.17
25	ITU	3	2.88	1.25	4	3.8	1.03	-1.6	3	2.88	1.25	3.5	3.62	1.06	-1.25
26	PU	3	3.12	0.99	3	3.6	0.84	-1.18	3	3.12	0.99	3	3.5	0.76	-0.99
27	PU	3.5	3.62	1.06	4	3.9	0.88	-0.66	3.5	3.62	1.06	4	3.88	0.99	-0.55
28	PU	2	2.5	0.76	3.5	3.4	0.7	-2.26 *	2	2.5	0.76	3.5	3.38	0.74	-2.06 ●
29	PU	3	2.88	0.83	3.5	3.4	0.7	-1.38	3	2.88	0.83	4	3.62	0.52	-1.87 ●
30	PU	3	3	0.76	4	3.7	0.82	-1.81 ●	3	3	0.76	4	3.5	0.76	-1.36
31	PU	3	3.38	0.52	4	3.9	0.57	-1.87 ●	3	3.38	0.52	4	3.75	0.46	-1.46
32	ITU	3	3.12	0.99	3	3.3	0.95	-0.47	3	3.12	0.99	3	3.25	0.71	-0.51

Note: ● - p-value <0.1, * - $p<0.05$, ** - $p<0.01$, *** - $p<0.001$.

TABLE VI: Statistics of the Results of the Post-task Questionnaire about the Catalogues

Q	Type	All participants							Good participants						
		GEN CAT			DOM CAT				GEN CAT			DOM CAT			
		Median	Mean	sd	Median	Mean	sd	Z_{MW}	Median	Mean	sd	Median	Mean	sd	Z_{MW}
1	ITU	3.5	3.5	1.2	4	3.3	0.95	0.38	3.5	3.5	1.2	3.5	3.12	0.99	0.66
2	ITU	3	2.88	1.25	3	2.9	1.1	0	3	2.88	1.25	2.5	2.62	1.06	0.44
3	PEOU	2.5	2.62	0.74	4	3.6	1.51	-1.77 ●	2.5	2.62	0.74	4	3.38	1.6	-1.29
4	PEOU	3	3	0.93	4	3.9	0.99	-1.81 ●	3	3	0.93	4	3.88	1.13	-1.57
5	PEOU	3	3	0.93	4	3.6	1.26	-1.31	3	3	0.93	4	3.5	1.41	-0.98
6	ITU	4	3.62	0.92	4	3.5	1.08	0.14	4	3.62	0.92	3.5	3.25	1.04	0.68
7	ITU	2	2.5	1.07	4	3.5	1.43	-1.62	2	2.5	1.07	4	3.25	1.49	-1.15
8	PEOU	3	3	0.76	3.5	3	1.49	-0.37	3	3	0.76	3	2.75	1.58	0.22
9	ITU	4	3.62	0.74	4	3.8	0.92	-0.46	4	3.62	0.74	4	3.75	1.04	-0.29
10	ITU	4	3.5	0.93	4	3.8	1.23	-0.78	4	3.5	0.93	4	3.75	1.39	-0.68
11	PU	3	3.38	0.92	4	3.6	0.84	-0.66	3	3.38	0.92	4	3.5	0.76	-0.51
12	PU	3.5	3.5	0.93	3	3.2	1.32	0.56	3.5	3.5	0.93	2	2.88	1.25	1.21
13	PU	4	3.75	1.04	4	3.8	1.03	-0.09	4	3.75	1.04	3.5	3.62	1.06	0.27
14	PU	3.5	3.38	1.06	4	3.7	1.06	-0.7	3.5	3.38	1.06	4	3.5	1.07	-0.28

Note: ● - p-value <0.1, * - $p<0.05$, ** - $p<0.01$, *** - $p<0.001$.

[9] T. Stålhane and G. Sindre, “Safety hazard identification by misuse cases: Experimental comparison of text and diagrams,” in *Proc. MODELS ’08*, 2008, pp. 721–735.

[10] T. Stålhane, G. Sindre, and L. Bousquet, “Comparing safety analysis based on sequence diagrams and textual use cases,” in *Proc. CAISE’10*, vol. 6051, 2010, pp. 165–179.

[11] P. Karpati, Y. Redda, A. L. Opdahl, and G. Sindre, “Comparing attack trees and misuse cases in an industrial setting,” *Inf. Soft. Technology*, vol. 56, no. 3, pp. 294 – 308, 2014.

[12] K. Labunets, F. Massacci, F. Paci, and L. M. Tran, “An experimental comparison of two risk-based security methods,” in *Proc. of ESEM ’13*, 2013, pp. 163–172.

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

[14] T. Stålhane and G. Sindre, “A comparison of two approaches to safety analysis based on use cases,” in *Proc. of ER ’07*, vol. 4801, 2007, pp. 423–437.

[15] K. Yskout, R. Scandariato, and W. Joosen, “Do security patterns really help designers?” in *Proc. of ICSE ’15*, 2015.

[16] UNITN, “Experiment page,” http://securitylab.disi.unitn.it/doku.php?id=winter-schl-exp2014.

[17] SESAR, *ATM Security Risk Assessment Methodology. SESAR WP16.2 ATM Security*, February 2003.

[18] BSI, *IT-Grundschutz Catalogues*, https://gsb.download.bva.bund.de/BSI/ITGSKEN/IT-GSK-13-EL-en-all_v940.pdf, 2005.

[19] F. D. Davis, “Perceived usefulness, perceived ease of use, and user acceptance of information technology,” *MIS Quarterly*, pp. 319–340, 1989.

[20] A. D. Re, *R Package ’compute.es’: Compute Effect Sizes*, 2014.

[21] R. Scherer, *R Package ’samplesize’: Sample size calculation for various t-Tests and Wilcoxon-Test*, 2012.

[22] C. Wohlin, P. Runeson, M. Hst, M. C. Ohlsson, B. Regnell, and A. Wessln, *Experimentation in software engineering*. Springer, 2012.