The Role of Catalogues of Threats and Security Controls in Security Risk Assessment: An Empirical Study with ATM Professionals

Martina de Gramatica[1], Katsiaryna Labunets[1(✉)], Fabio Massacci[1], Federica Paci[1], and Alessandra Tedeschi[2]

[1] DISI, University of Trento, Trento, Italy
{martina.degramatica,katsiaryna.labunets,fabio.massacci, federica.paci}@unitn.it
[2] Deep Blue srl, Roma, Italy
alessandra.tedeschi@dblue.it

Abstract. [**Context and motivation**] To remedy the lack of security expertise, industrial security risk assessment methods come with catalogues of threats and security controls. [**Question/problem**] We investigate in both qualitative and quantitative terms whether the use of catalogues of threats and security controls has an effect on the actual and perceived effectiveness of a security risk assessment method. In particular, we assessed the effect of using domain-specific versus domain-general catalogues on the actual and perceived efficacy of a security risk assessment method conducted by non-experts and compare it with the effect of running the same method by security experts but without catalogues.

[**Principal ideas/results**] The quantitative analysis shows that non-security experts who applied the method with catalogues identified threats and controls of the same quality of security experts without catalogues. The perceived ease of use was higher when participants used method without catalogues albeit only at 10 % significance level. The qualitative analysis indicates that security experts have different expectations from a catalogue than non-experts. Non-experts are mostly worried about the difficulty of navigating through the catalogue (the larger and less specific the worse it was) while expert users found it mostly useful to get a common terminology and a checklist that nothing was forgotten.

[**Contribution**] This paper sheds light on the important features of the catalogues and discuss how they contribute into risk assessment process.

Keywords: Empirical study · Security risk assessment methods · MEM

1 Introduction

Security risk assessment is a key step in the design of critical systems. Yet, system architects often lack the necessary security knowledge to identify all security risks. Even experts focus on those risks which according to their experience were

S.A. Fricker and K. Schneider (Eds.): REFSQ 2015, LNCS 9013, pp. 98–114, 2015.
DOI: 10.1007/978-3-319-16101-3_7

critical in the past. Thus, they can forget to treat risks which are less interesting for them, although they might be relevant for the system. To alleviate this issue, industrial security risk assessment methods and standards come with catalogues of threats and security controls. The catalogues can be divided by size and specialization into domain-general catalogues like BSI IT-Grundschutz Catalogues [3], ISO/IEC 27002 [8], NIST 800-53 [20], and domain-specific catalogues like PCI DSS [23] (Banking domain) or EUROCONTROL ATM [6] (Air Traffic Management domain).

In this paper we report an empirical study on the role of catalogues of threats and controls in conducting security risk assessment. The goal of the study is to assess the *actual and perceived efficacy of catalogues* in performing a security risk assessment by non-experts (with the catalogues) and by experts (using the same method but without catalogues). Actual effectiveness has been quantitatively investigated as the quality of threats and security controls identified by the participants. Perception has been assessed both quantitatively via post-task questionnaire and qualitatively via focus group interviews with the participants.

The study involved 15 professionals in the Air Traffic Management (ATM) domain who worked individually to identify threats and security controls for the Remotely Operated Tower (ROT) application scenario. More than two third of the participants had more than 5 years of experience in the ATM, while the others had at least 2 years of specific experience.

The main findings are that domain experts that are not security experts can obtain almost the same results as domain experts without catalogues while applying a security risk assessment method. Regarding perceived efficacy, domain-specific catalogues were perceived to be easier to use than domain-general ones because they are easier to navigate and there is a clear mapping between threats and security controls.

In addition, the analysis of focus group interviews shows that non-experts and security experts have a different perception of catalogues. Non-experts found catalogues useful as starting point to identify threats and controls but at the same time they were concerned about the difficulty in navigating the catalogues because there were no link between threats and security controls. Security experts instead found catalogues mostly useful because they provide a common terminology to discuss about threats and controls and they can be used to check completeness of results.

In the remainder of the paper, Section 2 presents the research method; Section 3 presents the motivation of domain selection and Section 4 describes the setting of the study, whose findings are presented in Sections 5 and 6. Threats to validity to our study are discussed in Section 7 and Section 8 presents the related work on prior research in the area. The findings and conclusion are presented in Section 9.

2 Research Method

The goal of this study is to investigate whether catalogues of threats and security controls facilitate the execution of a security risk assessment process. In particular, we want to assess whether the use of catalogues has an effect on the actual

and perceived efficacy of security risk assessment when used by people with no security expertise and comparing it with the effect of running the same assessment by security experts without catalogues. Accordingly, we formulated our research questions:

RQ1 *Does the use of domain specific or general catalogues improve the actual or perceived efficacy of a security risk assessment in comparison to each other and to the same assessment performed by experts without catalogues*?

RQ2 *Which are the qualitative features of a catalogue that impact actual or perceived efficacy*?

As our study is exploratory in nature, we applied a research approach combining both qualitative and quantitative methods. In particular, to address research questions *RQ*1 on actual and perceived efficacy we used a quantitative approach and divided the participants into three groups: the first group conducted a security risk assessment with the support of a domain-specific catalogue (DOM CAT), the second group with the support of a domain-general (GEN CAT) one, while the third group worked without catalogue (NO CAT). All participants in the NO CAT group had security knowledge, while most of the participants in the DOM CAT and GEN CAT groups had limited or none security knowledge.

Then, we measured *actual efficacy* as the quality of results produced by the participants. Two security experts independently assessed the quality. 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 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.

To measure *perceived efficacy* we asked the participants to fill in a post-task questionnaire along the Method Evaluation Model (MEM) [19]. According to MEM, we broke down perceived efficacy in *perceived ease of use (PEOU)* and *perceived usefulness (PU)*, and included the corresponding questions in the post-task questionnaire. The concrete post-task questions were adopted from the work of Opdahl and Sindre [21] in order to make comparison with related work easier. Questions were formulated as opposite statements with answers on a 5-point Likert scale. Table 3 in the appendix reports the post-task questionnaire.

To answer research question *RQ*2 we involved participants in focus group interviews where they answered questions on the process followed to identify threats and controls and their perception of the method and the catalogues. We investigated the transcripts of the interviews through the open coding methodology [29, Chap. 8], on the basis of a pre-defined set of codes, slightly edited from a list of codes used in previous studies [12,13]. This selection of codes allowed to identify the most frequently mentioned topics in the interviews. We considered these topics as the most representative in the discussion. The qualitative analysis attempted to cast light on the catalogues' features affecting actual and perceived effectiveness of security risk assessment.

3 Domain Selection

One of the key issues to conduct our study is the selection of an appropriate domain. The ATM domain has been often used in Requirements Engineering. For example, see the work of Maiden and Robertson [14] for general Requirements Engineering and our own for Security Requirements Engineering [16]. We also selected this domain because security plays an important role to ensure the resilience of ATM Service provision. To this end, the SESAR (Single European Sky ATM Research Program) project 16.02.03 focuses on analyzing existing approaches for security risks identification and tailoring them to the ATM domain.

The SESAR ATM Security Risk Assessment Method (SecRAM), developed within the project 16.02.03 [25], is the " official " method applied by ATM professionals in the SESAR program. SESAR designed SecRAM as a simple, step-wise method that should be applicable to all the SESAR Operational Focus Areas (OFAs). The overall 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. The method should be clear to personnel with little expertise and background in security and risk management. It is also should support the integration and comparison of security risk assessment results from different SESAR OFAs. In order to support non-expert, ATM professionals considered catalogues of threats and security controls as a great added value to carry out efficient and effective security risk assessment in SESAR.

We selected SecRAM as a reference security risk assessment method under study aiming to compare its effectiveness with domain-specific and domain-general catalogues. As instances of domain-specific and domain-general catalogues we selected EUROCONTROL ATM catalogues and BSI IT Grundschutz catalogues, respectively.

The ATM 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 describe 32 threats of three types: Physical, Information and Procedural. The catalogues also propose 33 pre and 18 post controls to mitigate each threat. Each control is linked to the mitigated threats and a description of the security control procedure.

The BSI IT-Grundschutz standard was developed by Bundesamt für Sicherheit in der Informationstechnik (BSI[1]), 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 describe 621 threats of the following types: Basic threats, Force Majeure, Organizational Shortcomings, Human

[1] Federal Office for Information Security (English).

Error, Technical Failure and Deliberate Acts. The safeguards catalogues describe 1444 security controls related to Infrastructure, Organization, Personnel, Hardware and software, Communication and Contingency planning.

The application scenario was chosen among one of the ATM new operational scenarios that have already been assessed by SESAR with the SecRAM methodology: the Remotely Operated Tower (ROT).

The Remote and Virtual Tower, is a new operational concept proposed by SESAR [26,27]. The main change with respect to current operations is that control tower operators will no longer be located at the aerodrome. They will move to a Remotely Operated Tower Center. Each tower module will be remotely connected to (at least) one airport and consist of one or several Controller Working Positions. The operator will be able to do all air traffic management tasks (e.g. authorize landing, departure, etc.) from this position. The idea is that operator will be able to control remotely more than one airport. The visual surveillance will be provided by a reproduction of the Out of The Window view, by using visual information capture and/or other sensors such as cameras with a 360-degree view, which will be able to zoom 36 times closer than current binoculars in all weather conditions. The visual reproduction can be overlaid with information from additional sources if available, for example, surface movement radar, surveillance radar, or other positioning and surveillance implementations providing the positions of moving object within the airport movement area and vicinity. The collected data, either from a single source or combined, is reproduced for the operator on data/monitor screens, projectors or similar technical solutions. The use of technologies will also enhance the visual reproduction in all visibility conditions (e.g., bad weather conditions).

This scenario presents relevant ATM and security issues and technological challenges that can benefit from a Security Risk Assessment. As apparent from the description, the ROT concept will be encompassed by data confidentiality, integrity and availability issues, also affecting airport safety, as well as physical security issues, like the on-site protection of the remotely located cameras, sensors and surveillance radars in the aerodrome, to be analyzed during our experiment.

4 Execution and Demographics

The study was run in May 2014 at Deep Blue premises and consisted of an empirical study with 15 professionals from several ATM Italian companies. As mentioned before the participants were divided into three groups and assigned to three different treatments. They were asked to apply individually the same method, namely SESAR SecRAM, with the support of domain-specific catalogues (EUROCONTROL ATM), general-domain catalogues (BSI IT-Grundschutz) or without any catalogues. Before starting, the participants were administered a questionnaire to collect information on their background and previous knowledge of other risk assessment methods.

The study was based on a step-wise process consisting of three interacting phases:

Table 1. Participants' Demographic Statistics

Variable	Scale	Mean	Distribution
Age	Years	33.1	20% were 25-29 years old; 53.3% were 30-39 years old; 20% were 40 and older
Gender	Sex		66.7% male; 33.3% female
Academic Degree			73.3% had MSc degree; 26.7% had PhD degree
Work Experience	Years	7.9	26.7% had $\geq$ 2 and <5 years; 46.7% had $\geq$ 5 and <10 years; 26.7% had $\geq$ 10 years
Experience in Risk Assessment	Years	0.67	Three participants had 2 years, 1.5 years and 0.25 years, respectively
Security/Privacy Knowledge	Yes/No	-	47% had experience; 53% had no experience

Training. The application scenario description was administered to participants for an individual reading. A frontal-training phase followed in which the method designer introduced the considered methodology process through a step by step tutorial.

Application. Each step of the method introduced in the tutorial, was forthwith applied individually on the case study until the completion of the last step.

Evaluation. Three evaluators independently judged the quality of the threats and security controls identified by the participants, providing marks and comments.

After the application phase we administered to the participants a post-task questionnaire to gather their perception of the method and the catalogues employed. They were later involved into focus groups, according to their treatment, to discuss drawbacks and benefits of the method and the catalogues under study. A list of questions guide the discussion that had been audio recorded for further analysis. The main positive and negative aspects generated in the focus groups then were reported on post-it notes.

The participants of the study were 15 practitioners from the different ATM companies. Table 1 presents descriptive statistics about the participants. Most of the participants (73.4%) reported that they had at least 5 years of working experience, some participants (26.7%) reported from 2 to 5 years of workings experience. In addition, almost half of participants (47%) reported that they had security/privacy knowledge, the rest did not report any similar knowledge. Three out of sixteen participants reported from 3 months up to 2 years experience in security risk assessment.

5 Quantitative Results

In this section we discuss the results on actual efficacy of the risk assessment and perceived efficacy of the method and catalogues. Tables 2a and 2b report the median values for Actual Efficacy, PU and PEOU of the method for each treatment. The detailed results of risk assessment delivered by the participants are reported in the Table 4. The detailed statistics on post-task questionnaire responses are reported in Table 5.

Table 2. Summary of Quantitative Results

(a) Threats

	DOM CAT	GEN CAT	NO CAT
AE	3.5	2.5	2.5
PU	4	4	4
PEOU	3	4	4

(b) Security Controls

	DOM CAT	GEN CAT	NO CAT
AE	3.5	2.5	3
PU	4	4	4
PEOU	3	3	4

The AE row reports the medians of experts assessment of the threats and security controls produced by the participants. The PU (respectively PEOU) row reports the medians of participants' responses to a post-task questions about method's PU (PEOU). All values are on a 1-5 scale with 5 being the best score. The columns describe the type of task performed by the participants: risk assessment with a domain specific catalogue (DOM CAT), a generic catalogue (GEN CAT), or no catalogue by security experts (NOCAT).

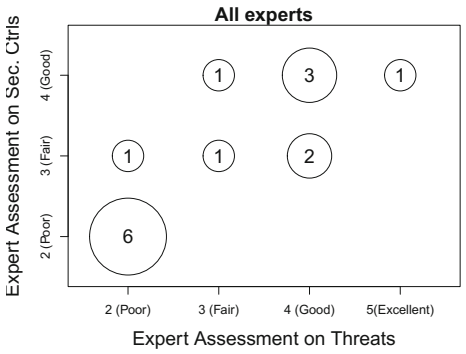

Fig. 1. Experts assessment of quality of threats and security controls

Actual Efficacy. As mentioned before we measured method's actual efficacy as a quality of threats and security controls identified by the participants. Two ATM security experts independently assessed the quality. They are reported a similar assessment for each group. Figure 1 illustrates the average of experts' evaluation for threats (reported on x-axis) and security controls (on y-axis). Six participants out of fifteen performed poorly. In terms of the final assessment we observed that: *a)* the experts marked bad participants the same way; *b)* they consistently marked moderately good participants; and *c)* they had a different evaluation only for the threats of one participant and for the security controls of another participant out of 15 participants.

We used Wilcoxon test to validate if the difference in experts' evaluation is statistically significant. The results showed that there is no statistically significant differences in the evaluations of two experts both for threats ($p = 0.09$) and controls ($p = 0.77$).

The first lines in Tables 2a and 2b report the quality of threats and security controls identified with three treatments. We used Kruskal-Wallis (KW) test to investigate the statically significant difference in the quality between treatments.

Table 2a shows that participants who used domain-specific catalogue to identify security controls performed as participants who did not use the catalogues.

While, the participants who applied the domain-general catalogue performed even worst than participants without catalogue. The difference in the quality of security controls is not statistically significant based on the results of KW test. Therefore we can conclude that there is no difference in the actual efficacy of a security risk assessment when used with catalogues by non-experts and without catalogues by security experts.

Perceived Efficacy. Table 2a shows that there is no difference in method's PU when the method is applied with or without catalogues of threats. Same results we have for method's PU regarding security controls identification (see Table 2b). Considering method's PEOU, the participants who conducted threats identification with domain-general catalogue of threats or without catalogue reported higher method's PEOU than participants who applied the domain-specific catalogues. While for method's PEOU for security controls identification only the participants who conducted risk assessment without catalogues reported higher perception. We also used non-parametric KW test to analyze the differences in participants PU and PEOU of the method. However, the results of KW test did not reveal any significant differences in PU and PEOU except one. The results of KW test showed: there is 10% significant difference in method's PEOU with respect to security controls identification (KW p =0.099). However, the post-hoc analysis with Mann-Whitney test with Holm correction [10, Chap. 14.2] did not show any significant differences between treatments. Therefore, we can conclude that there is no difference in the perceived efficacy of the method when used by non-experts with catalogues and by security experts without catalogues.

Exploring Correlations. We also explored possible correlations between actual and perceived efficacy with Kendall tau rank correlation coefficient. We used this test because our data are ordinal and have many ties. The correlation test revealed only one significant relation between the quality of threats and participants' answers to the question "*method helped me in brainstorming on the security controls*". This is positive statistically significant correlation (p =0.04, $\tau = 0.45$).

6 Qualitative Results

In this section we report the analysis result of focus groups interviews and post-it notes sessions with the participants. The results explain the differences in the perception of two types of catalogues and outline the key features that effective catalogues must have.

Catalogue Structure. The analysis of interviews shows that the structure of catalogue is a key aspect in the identification of threats and security controls. Thanks to its basic layout, the clear tables and its relative length, the domain-specific catalogue is generally perceived by the participants as easier to browse and to read: "*I read only the titles* [namely the reference to the "Generic Threat" and the "Attack Threat"], *they were quite explanatory, therefore a very short*

consultation of the catalog allowed me to produce enough content" [DOM CAT participant]. This is particularly true in comparison with the domain-general catalogue, consisting of a long list of items, perceived as "*not user-friendly at a first read*" [GEN CAT participant] and "*difficult to navigate and master due to its length and structure*" [GEN CAT participant].

Another relevant aspect in the structure of the domain-specific catalogues is the presence of linking references between threats and security controls. According to some participants this feature makes the identification of the controls an automatic mechanism: "*Once identified the threat, finding out controls was really a mechanical work*"[DOM CAT participant]. Even more so for security-novices, traceability is perceived as a fundamental feature in the structure of the catalogue. Because it provides a one-directional link between the two objects of interest, that makes the mistake quite impossible. In contrast, the domain-general catalogue does not provide this support and therefore the findings are affected: "*The identification of security controls was more difficult because you had to map them with the threats previously identified but there was no direct link in the catalogue. It was mainly due to a problem of usability of the catalogue*" [GEN CAT participant]. Examples, present in the specific-domain catalogue, are also perceived as helpful in the identification of threats and security controls.

Based on these findings we can conclude that a series of paths through structure of the catalogue will facilitate the threats and security controls identification. Thus, the usability of the catalogues is of capital importance mostly for security non-experts. The same we can said about navigability and traceability, two of the features that make the domain-specific catalogue a practical and useful tool for the risk assessment.

Catalogue Size and Coverage. If a catalogue is meant for security-novices the abstraction level should be kept low and just provide few critical threats and security controls. Otherwise, the security-novices can feel overwhelmed and not able to find any threat or security control at all. This is particularly the case of the general-domain catalogue, judged as: "*Very difficult to consult for non-technical people*"[GEN CAT participant] given the high number of threats and controls proposed. An interesting statement in this regard, comes from a participant who was not assigned to any catalogue but had the chance to glance at the general-domain catalogue. His opinion expresses the potential problem inherent to the use of a too complex catalogue: "*I saw people near to me; they were not able to find out stuff in the catalogue, they kept on getting lost in the pages and eventually they came up always with the same two or three items*"[NOCAT participant].

Regarding the coverage instead, considered in terms of specificity of the items , the opinion expressed by the participants was quite contrasting: this is particularly proven by the statements from the security experts claiming that the suggestions in both catalogues were very generic, rather than specific, precise and well-defined threats and controls: "[The catalogue provided a] *list of non-specific threats impacting the specific concept under investigation*" [GEN CAT participant] (from a security-expert user). The same result comes from the domain-

specific catalogue: "*I found the catalogue useful, but I noticed that many threats were repeated*" [DOM CAT participant]. While security-novices did not support the idea and seems were in general more satisfied by the use of the catalogue. This is probably due to the fact that, without any experience any kind support is of great benefit. Security-novices than could not be able to judge the quality of the results achieved given their little past experience.

To be a useful tool for security experts the catalogue must provide specific threats and controls, otherwise it only allows to define generic and thus ineffective controls.

Catalogue as Common Language. One feature of the catalogue perceived as essential by every participant, irrespectively of the type of catalogue employed, is the fact that a catalogue by itself provides a common terminology for all users. As suggested by one participant, "*The catalogue could be seen as a useful tool, able to formalize the controls that have been formulated in an informal way, and to lead them back into a common nomenclature*" [DOM CAT participant]. "*The problem arises when we are in the same group and we use a different language*" [NOCAT participant]. The demand for a standard language caused by the need of sharing, discussing and presenting results that could be understood and therefore adopted by all participants of the risk assessment process. Unsurprisingly, this aspect is mostly perceived as important by participants who were not assigned to any catalogue.

Catalogue as Check-list. One tendency identified in the analysis is the difference in the opinion of security experts and security-novices about their general perception towards the catalogue. Security-novices indeed are more prone to express a positive judgment on the benefit of using the catalogue. While security-experts tend to be more uncertain about the real advantages of the catalogue. This could be explained by the fact that the catalogue represents an essential support for users without any (or with little) experience, as claimed here: "*The catalogue is really helpful if you do not have any background*" [DOM CAT participant]. While the added value for experienced users is not as higher as expected.

Furthermore, the statements collected from security-experts suggest an additional aspect: "*The first step is to use your own experience and then to use the catalogue to cover generic aspects that could be forgotten*"[NOCAT participant]. For security-experts the catalogue is perceived as a check-list, as something that can be used after a brainstorming session where user works based on his own experience. In this way, the catalogue is supposed to provide the verification of the efficiency and the coverage of the threats and security controls identified. For security-novices on the contrary, the catalogue represents: "*A good starting point for the evaluation of the threats and the controls.*" [DOM CAT participant].

Catalogue and Knowledge. Participants with security knowledge cared more about the quality of threats and security controls that they could identify with the support of the catalogues. That is mainly due to the fact that they used their expertise to evaluate the achieved results. Security experts based on their

previous knowledge expected more specific results from the support of the catalogue. While security-novice were not able to judge the quality of the identified threats and controls. Therefore, they were more concerned about the usability of the catalogues, as demonstrated by their observations on the traceability and the navigability of the catalogues (see sections above).

7 Threats to Validity

The main threats to validity are related to internal, conclusion and external validity [30].

Another threat to internal validity could be the size of catalogues as the domain-general catalogues are significantly lager than the domain-specific ones in order to cover more grounds. We mitigated this threat by making the use of domain-general catalogues of similar difficulty as domain-specific one (155 pages) we prepare a short version of 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 (˜2500 pages).

The main threat to conclusion validity is related to the *sample size* that must be big enough to come to correct conclusions. We aware that due to the low number of participants (N=5 × 3) it is unlikely to draw any strong statistical results. But Meyer et al. [18] show that it is possible to have statistically significant results for the samples contain 3 and more observations. To control possible effect of participants' background on the results we collect information about participants' through demographics and background questionnaire at the beginning of the study. To mitigate possible effect of previous knowledge about object of the study the participants were given a step by step tutorial on the security risk assessment method and received textual description of the application scenario.

Another threat to conclusion validity could be the number of security risk assessment which produced low quality threats and controls based on the experts evaluation (6 out of 15). However, we think the level of quality reflects the diversity of participants' knowledge and expertise. It could be a threat to validity if we would have had all the risk assessment producing threats and controls of the same quality.

The main threat to external validity external validity is that both the risk assessment method and scenario were chosen within the ATM domain. However, the chosen risk assessment method is compliant with ISO 27005 standard that can be applied to different domains not just to the ATM. Therefore, this threat is not present in our study.

8 Related Work

In this section we reviewed the studies that relevant to our work that are studies comparing security methods and studies which investigated the role of structured knowledge in Requirement Engineering (RE).

Empirical Evaluation of Security Methods. There are many catalogues that describes existing security problems and countermeasure. We can divide them into general catalogues that describe Information Systems security practices like BSI IT-Grundschutz Catalogues [3], ISO/IEC 27002 and 27005 [7,8], NIST 800-30 and 800-53 [20,28], COBIT 5 [1,4], or domain-specific catalogues like PCI DSS [23] for banking security, or EATM for security and safety in ATM, OWASP [22] for web application security.

Yet, most of the studies evaluate the effectiveness of the risk assessment process detached from the security knowledge [9,11,12,21,24]. The effect of the use of catalogues on the actual and perceived effectiveness of risk assessment is not yet studied. And it is still a question which catalogues' aspects affect actual effectiveness of risk assessment and how they impact user perception.

Opdahl and Sindre [21] reported two controlled experiment with 28 and 35 students to compare attack trees and misuse cases. In [11] the same group of researchers reported the replication of the experiment with industrial professionals. Both experiments showed that attack trees help to identify more threats than misuse cases. In our study we adopted similar perception variables and post-task questions to measure them.

Jung et al. [9] reported two controlled experiments (7 PhD students and 11 practitioners) to compare two safety analysis methods, namely Fault Trees (FT) and Component Integrated Fault Trees (CFT). The methods were compared with respect to the quality of the results and participants' perception. The experiments showed that CFT could be beneficial for users without expertise in FT. Similar to this work, we adopted quality of results as a way to measure actual effectiveness of the method.

Among the experiments which studied industrial security assessment methodologies, Scandariato et al. [24] reported a descriptive study with 41 MSc students to observe how STRIDE works in laboratory conditions. The goal of this study was to assess STRIDE with respect to productivity of participants, and the correctness and completeness of the results. The participants were trained on STRIDE application during three lectures that is a reasonable time for training. As an application scenario was chosen a medium-scale distributed Digital Publishing System. The participants had 4 hours to apply STRIDE in the class and were allowed to finish the task as homework. The results of the experiment showed that precision of the results was acceptable but their productivity was quite low. In our study we selected a mix-method approach to evaluate both performance of the participants and their perception of risk assessment method and catalogues. We also completed our study with focus groups interview and post-it notes session in order to investigate the reasons behind quantitative results and shed light on the corresponding specific aspects of catalogues.

Labunets et al. [12] reported controlled experiment with 28 MSc students to compare the actual effectiveness and perception of visual (CORAS) and textual (SREP) methods for security risk assessment. The results of the experiment showed that visual method is more effective in identifying threats and better perceived by the participants than the textual one. Similar to previous study, the

recent work of Labunets et al. [13] reported controlled experiment with 29 MSc students to compare textual (EUROCONTROL SecRAM) vs. visual (CORAS) industrial security risk assessment methods. The results showed that there is no difference in actual effectiveness of two methods, but the visual method had better perception. In our study we adopted similar experimental protocol proposed in [13]. We also adopted similar dependent variables (actual effectiveness, perceived usefulness and perceived ease of use). It is noteworthy to mention that in [13] participants reported that security risk assessment methods "*would benefit from availability of catalogues of threats and security controls*".

Considering similar empirical studies in the ATM domain it is worthy to mention the works of Maiden et al. [14,15]. They reported several case studies in ATM domain to evaluate the effectiveness of RESCUE, a scenario-driven requirements engineering method. The studies were conducted as series of RESCUE workshops with ATM professionals from different backgrounds. The participants applied method to gather requirements for the real complex ATM systems. The authors collected qualitative data by mean of post-it notes, color-coded idea cards and pin boards. The results of the studies demonstrated the effectiveness of the RESCUE method. Similar to Maiden et al., we conducted our study in form of two-days workshop with ATM professionals from different backgrounds. We concluded workshop with focus group interviews with participants to collect their opinion about most important aspects of the catalogues.

Empirical Studies on the Role of Structured Knowledge. The role of structured knowledge, i.e. catalogues, has not been investigated in the security community, but it has been investigated in RE community.

The work of Mavin and Maiden [17] is the closest to our study. This work aimed to investigate if structured knowledge have an effect on the effectiveness of walkthrough techniques and, therefore, led to better effectiveness in elicitation of stakeholder requirements. They also investigate if the domain-specific scenarios increase the effectiveness of requirements elicitation comparing to the other technique. The authors conducted a case study with a team of ATM professionals. The results showed that the use of walkthroughs with domain-specific scenarios doubled the number of elicited requirements comparing to the other method that was used by the team over the previous 6 months. In our study we also aimed to investigate the effect of knowledge on the effectiveness of the security risk assessment. In our case knowledge introduced into security risk assessment process in form of domain-specific or domain-general catalogues of threats and security controls.

To the best of our knowledge there is only one study aiming to investigate the effectiveness of using catalogues but in requirements engineering. Cysneiros [5] evaluated the effectiveness of using catalogues on nonfunctional requirements elicitation. The paper reported a controlled experiment with 12 fourth year students. The results of the experiment showed that the groups used catalogues with a method performed better than the others participants applied either method without catalogues or catalogues without method. However, there is no similar papers aiming to investigate effectiveness of catalogues of threats and

security controls. In our study we compared the effect of using domain-specific and domain-general catalogues vs. using just security risk assessment method on the actual and perceived effectiveness.

9 Discussion and Conclusions

Security catalogue is an important part of security risk assessment process. Barnum and McGraw [2] admitted a crucial role of catalogues: *"as the [security] field evolves and establishes best practices, knowledge management can play a central role in encapsulating and spreading the emerging discipline more efficiently."*

The aim of catalogues of threats and security controls is to put best security practices into uniform document that can be re-used in security risk assessment. In this paper we have investigated in both qualitative and quantitative terms the effect of using domain-specific catalogues versus domain-general catalogues, and compare them with the effects of using the same method by security expert but without catalogues.

In quantitative terms there is no difference in the actual effectiveness of a security risk assessment method when used with catalogues by non-experts and without catalogues by security experts, albeit only few groups achieved a high quality score in terms of identified threats and security controls.

The qualitative analysis, carried with focus group interview and post-it notes session, showed that security experts have a different expectations from a catalogue than non-experts. Non-experts were mostly worried about the difficulty of navigating through the catalogue while expert users found it mostly useful to get a common terminology and a checklist that nothing was forgotten.

The catalogue alone does not facilitate the identification of threats and security controls. Participants without security knowledge were able to identify some threats and controls but these were not specific for the scenario under analysis. Participants who used the catalogues and had security knowledge were able to produce good threats and controls. Those who had security knowledge and did not use any catalogue performed the same or sometimes even worse than other participants. Catalogues could provide support for discussion among the analysts because they provide a common language for analysts with different background. They could also be used to check the completeness and coverage of the results.

In summary, the study show that with the use of the catalogues a satisfactory number of threats and controls can be identified. Results of higher quality can be better achieved through a combination of the catalogue and the added value of experience. If the latter is expensive to get, a domain-specific catalogue is your second best bet.

Acknowledgments. This work has been partly supported by the EU under grant agreement n.285223 (SECONOMICS) and by the SESAR JU WPE under contract 12-120610-C12 (EMFASE).

A Additional information

Table 3. Post-task Questionnaire

Q#	Type	Question (positive statement)
1	PEOU	SecRAM helped me in brainstorming on the threats
2	PEOU	SecRAM helped me in brainstorming on the security controls
3	PEOU	I found SecRAM easy to use
4	PU	SecRAM process is well detailed
5	PEOU	SecRAM was difficult to master
6	PEOU	I was never confused about how to apply SecRAM to the application
7	PU	I would have found specific threats more quickly with the SecRAM
8	PU	I would have found specific security controls more quickly with the SecRAM
9	PU	SecRAM made the security analysis more systematic
10	PEOU	SecRAM made it easier to evaluate whether threats were appropriate to the context
11	PEOU	SecRAM made it easier to evaluate whether security controls were appropriate to the context
12	PU	SecRAM made the search for specific threats more systematic
13	PU	SecRAM made the search for specific security controls more systematic
14	PU	If I need to update the analysis it will be easier with SecRAM than with common sense
15	PU	SecRAM made the security analysis easier than an ad hoc approach
16	PU	SecRAM made me more productive in finding threats
17	PU	SecRAM made me more productive in finding security controls

Table reports post-task questions and their perception type, PU or PEOU (questions about intention to use and perceive leverage are omitted). Some questions do no specify whether the method was used for threats or for controls. In that case we have used the corresponding answers for both threats and controls.

Table 4. Participants, Their Results and Quality Assessment

ID	Security	Working	Education	Catalog	Quantity		Quality (Exp1)		Quality (Exp2)	
	Knowledge	Experience	Length		Threats	SecCtrls	Threats	SecCtrls	Threats	SecCtrls
P01	No	6	MSC	GEN CAT	17	28	2	2	3	3
P02	No	5	PHD	GEN CAT	9	17	1	2	2	2
P03	Yes	4	MSC	GEN CAT	27	50	4	4	4	3
P04	No	5	MSC	GEN CAT	9	23	2	2	3	3
P05	Yes	4	PHD	GEN CAT	9	15	3	3	3	3
P06	No	8	DIPLOMA	DOM CAT	22	38	4	3	3	3
P07	No	4	MSC	DOM CAT	7	14	2	2	2	2
P08	No	5	PHD	DOM CAT	24	66	4	4	4	4
P09	Yes	2	MSC	DOM CAT	24	45	5	4	5	4
P10	No	7	PHD	DOM CAT	16	32	4	4	3	3
P11	No	5	MSC	NOCAT	10	13	2	1	3	3
P12	Yes	14	PHD	NOCAT	15	47	3	3	4	3
P13	Yes	17	MSC	NOCAT	15	19	2	3	3	3
P14	Yes	18	MSC	NOCAT	24	28	2	2	3	3
P15	Yes	15	MSC	NOCAT	6	13	2	4	4	3

Table presents the information about security knowledge, working experience and degree of participants; number of threats and security controls identified by participants and the assessment from two ATM experts on the quality of threats and security controls.

Table 5. Responses to the Post-task Questions

Q#	Type	DOM CAT		GEN CAT		NO CAT	
		Mean	Median	Mean	Median	Mean	Median
1	PEOU	4.2	4	4	4	3.2	3
2	PEOU	4.2	4	3.2	4	3.2	3
3	PEOU	3.4	3	3.2	4	4.2	4
4	PU	3.4	4	3.4	3	3.8	4
5	PEOU	3	3	3.4	4	3.8	4
6	PEOU	2.8	3	2.6	3	4	4
7	PU	3.4	3	2.4	2	3.2	3
8	PU	3.8	4	2.4	2	3.2	3
9	PU	3.8	4	4.2	4	4.2	5
10	PEOU	3.2	3	3.4	4	3	3
11	PEOU	2.8	3	2.6	2	3	3
12	PU	3.8	4	3.8	4	3.6	3
13	PU	3.4	3	3.6	4	3.6	4
14	PU	4	4	3.6	4	4.6	5
15	PU	2.8	3	2.6	3	3.6	4
16	PU	4.2	4	3	4	3.4	4
17	PU	4	4	3.4	4	3.4	3

Table reports mean and median value of participants' responses to each post-task question and the type of the question.

References

1. Information System Audit and Control Association: COBIT 5: A Business Framework for the Governance and Management of Enterprise IT (2012)
2. Barnum, S., McGraw, G.: Knowledge for software security. IEEE Security & Privacy **3**(2), 74–78 (2005)
3. BSI: IT-Grundschutz Catalogues (2005)
4. COBIT: Control Practices: Guidance to Achieve Control Objective for Successful IT Governance, 2nd edn. IT Governance Institute (2007)
5. Cysneiros, L.M.: Evaluating the effectiveness of using catalogues to elicit non-functional requirements. In: WER, pp. 107–115 (2007)
6. EATM: Threats, pre-controls and post-controls catalogues. European Organisation for the Safety of Air Navigation (2009)
7. ISO: Iso/iec 27005: Information technology security techniques - information security risk management (2012)
8. ISO: IEC 27002: 2013 (EN) Information technology-Security techniques-Code of practice for information security controls Switzerland. ISO/IEC (2013)
9. Jung, J., Hoefig, K., Domis, D., Jedlitschka, A., Hiller, M.: Experimental comparison of two safety analysis methods and its replication. In: 2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 223–232. IEEE (2013)
10. Juristo, N., Moreno, A.M.: Basics of software engineering experimentation. Springer Publishing Company, Incorporated (2010)
11. Karpati, P., Redda, Y., Opdahl, A.L., Sindre, G.: Comparing attack trees and misuse cases in an industrial setting. Inf. Soft. Technology **56**(3), 294–308 (2014)
12. Labunets, K., Massacci, F., Paci, F., Tran, L.M.: An experimental comparison of two risk-based security methods. In: Proc. of ESEM 2013, pp. 163–172 (2013)

13. Labunets, K., Paci, F., Massacci, F., Ruprai, R.: An experiment on comparing textual vs. visual industrial methods for security risk assessment. In: 2014 IEEE Fourth International Workshop on Empirical Requirements Engineering (EmpiRE), pp. 28–35. IEEE (2014)
14. Maiden, N., Robertson, S.: Integrating creativity into requirements processes: experiences with an air traffic management system. In: Proceedings of the 13th IEEE International Conference on Requirements Engineering, pp. 105–114. IEEE (2005)
15. Maiden, N.A.M., Jones, S.V., Manning, S., Greenwood, J., Renou, L.: Model-driven requirements engineering: synchronising models in an air traffic management case study. In: Persson, A., Stirna, J. (eds.) CAiSE 2004. LNCS, vol. 3084, pp. 368–383. Springer, Heidelberg (2004)
16. Massacci, F., Paci, F., Tran, L.M.S., Tedeschi, A.: Assessing a requirements evolution approach: Empirical studies in the air traffic management domain. Journal of Systems and Software (2013)
17. Mavin, A., Maiden, N.: Determining socio-technical systems requirements: experiences with generating and walking through scenarios. In: Proceedings of the 11th IEEE International on Requirements Engineering Conference, pp. 213–222. IEEE (2003)
18. Meyer, J.P., Seaman, M.A.: A comparison of the exact kruskal-wallis distribution to asymptotic approximations for all sample sizes up to 105. The Journal of Experimental Education **81**(2), 139–156 (2013)
19. Moody, D.L.: The method evaluation model: a theoretical model for validating information systems design methods. In: Proceedings of the 11th European Conference of Information Systems (ECIS), pp. 1327–1336 (2003)
20. NIST: SP. 800–53. Recommended Security Controls for Federal Information Systems, 800-53 (2013)
21. Opdahl, A.L., Sindre, G.: Experimental comparison of attack trees and misuse cases for security threat identification. Inf. Soft. Technology **51**(5), 916–932 (2009)
22. OWASP: The Ten Most Critical Web Application Security Risks 2013. The Open Web Application Security Project (2013)
23. PCI DSS: Payment Card Industry Data Security Standards. http://www.pcisecuritystandards.org
24. Scandariato, R., Wuyts, K., Joosen, W.: A descriptive study of microsoft's threat modeling technique. REJ, pp. 1–18 (2014)
25. SESAR: ATM Security Risk Assessment Methodology. SESAR WP16.02.03: ATM Security, February 2003
26. SESAR: Single Remote Tower Technical Specification Remotely Operated Tower Multiple Controlled Airports with Integrated Working Position - project P12.04.07 (2012)
27. SESAR: OSED for Remote Provision of ATS to Aerodromes - project P06.09.03 (2013)
28. Stoneburner, G., Goguen, A., Feringa, A.: Risk management guide for information technology systems. NIST special publication, 800-30 (2002)
29. Strauss, A., Corbin, J.M.: Basics of qualitative research: Grounded theory procedures and techniques. Sage Publications, Inc (1990)
30. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in software engineering. Springer (2012)