On the Equivalence Between Graphical and Tabular Representations for Security Risk Assessment

Katsiaryna Labunets[1], Fabio Massacci[1] and Federica Paci[2]

[1] DISI, University of Trento, Italy,
{name.surname}@unitn.it
[2] ECS, University of Southampton, UK
F.M.Paci@soton.ac.uk

Abstract. **[Context]** Many security risk assessment methods are proposed both in academia (typically with a graphical notation) and industry (typically with a tabular notation). **[Question]** We compare methods based on those two notations with respect to their actual and perceived efficacy when both groups are equipped with a domain-specific security catalogue (as typically available in industry risk assessments).
[Results] Two controlled experiments with MSc students in computer science show that tabular and graphical methods are (statistically) *equivalent in quality* of identified threats and security controls. In the first experiment the perceived efficacy of tabular method was slightly better than the graphical one, and in the second experiment two methods are perceived as equivalent. **[Contribution]** A graphical notation does not warrant by itself better (security) requirements elicitation than a tabular notation in terms of the quality of actually identified requirements.

Keywords: security risk assessment method; empirical study; controlled experiment; method evaluation model; equivalence testing

1 Introduction

Risk analysis is an essential step to deliver secure software systems. It is used to identify security requirements, to look for flaws in the software architecture that would allow attacks to succeed, and to prioritize tests during test execution.

Problem. An interesting observation is that there is a difference in notation between academic proposals and industry standards for security risk assessment (SRA). Most academic approaches suggest a graphical notation, starting from the seminal work on Anti-Goals [1] to [2] and more recently [3]. Industry opts for tabular models like OCTAVE [4], ISO 27005 and NIST 800-30. Microsoft STRIDE [5] is the exception on the industry side and SREP [6] is the exception on the academic side.

The initial goal of our long term experimental plan in 2011 [7] was to empirically prove that (academic) SRA methods using a graphical notation (for short

"graphical methods") were indeed superior to risk assessment methods using a tabular notation (for short "tabular methods"). We struggled to prove difference in our previous experiments [8,9], then maybe we should prove equivalence. Thus, our study aims to answer the following research questions (RQs):

RQ1 Are tabular and graphical SRA methods equivalent w.r.t. actual efficacy?
RQ2 Are tabular and graphical SRA methods equivalent w.r.t. perceived efficacy?

Approach. We ran two controlled experiments with 35 and 48 MSc students who worked in groups of two participants. They applied both methods to four different security tasks (i.e. 2 tasks per each method) for a large scale assessment lasting 8 weeks. In the first experiment groups analyzed security tasks for the Remotely Operated Tower (ROT) for Air Traffic Management (ATM). To prevent plagiarism between two experiments, in the second experiment we asked groups to perform the same security tasks but for a different ATM scenario, namely Unmanned Aerial System Traffic Management (UTM).

We measured *actual efficacy* as the quality of threats and security controls identified with a method as rated by domain-experts. *Perceived efficacy* is measured in terms of *perceived ease of use* (PEOU) and *perceived usefulness* (PU) of the methods through a post-task questionnaire administered to participants. The independent variables were methods and security tasks to assess.

A key difference with our previous studies (e.g. [10]) is that we provided to both groups a industry catalogue with hundreds of domain-specific threats and security controls. In this setting, using the number of identified threats and control as a measure of quality (as we did in our first study [10]) would have been inappropriate as anybody could obtain a large number of (potentially irrelevant) threats or controls just by looking up into the catalogue. So we employed several domain security experts to rate the result of the students.

We also replaced the academic tabular method SREP [6] which we used in [10] by a method used in the industry SecRAM [11] which had very similar tables but a nimbler process, designed by risk-assessment industry experts to simplify the risk assessment process, in the same fashion that the graphical method was designed by SINTEF to be simple to use in its industry consultancies [3].

Key Findings and Contribution. Our main findings — as unpalatable as they might be — are that, given the same conditions, the *tabular and graphical methods are equivalent* to each other with respect to the actual and perceived efficacy. Both results are *statistically significant* when compared with two one-sided tests (TOST) [12,13] which allows for testing for equivalence of outcomes.

Our study shows that representation by itself is not enough to warrant the superiority of a graphical model over a tabular model. Translating this result to general requirements engineering would mean that the fancy graphics of i* [14] and its many offsprings, are equivalent to the plain tables of Volere [15].

2 Background and Related Work

From an academic perspective, we have seen a significant development in requirements engineering towards graphical methods to identify security requirements. Some were backed up by formal reasoning capabilities [1,2], others offered variants of graphical notation [16,17,18,3], or minimal model based transformation analysis [19]. An epiphenomena of this trend was the RE'15 most influential paper award to the RE'05 paper introducing a graphical notation and sophisticated reasoning capabilities to verify security properties [2].

In contrast, industry standard development bodies doggedly use tabular representations for the elicitation of threats and security requirements. NIST 800-30 and ISO 27005 standards both use tables. Domain specific methodologies such as SecRAM [11], designed for risk assessment in ATM, also use tables. Most of tables use essentially the same wordings, with major differences being mostly on the process (some suggesting to analyze threats first, others suggesting to start the analysis from assets). Such preference could be due to simplicity, or the need to produce the documentation (in forms of table) that is often need to achieve compliance (as opposed to actual security).

As mentioned, our research goal since 2011 [7] has been to prove that graphical methods were actually superior to tabular methods. In all our experiments, in order to make the comparison fair, the difference between the methods was purely in the notation and the accompanying modeling process: graphical notation on one side, tabular on the other side. The formal reasoning capabilities supported by some methods [2] were never called into play.

This was never considered to be a problem, as the RE trend since 2005 has "revealed the emergence of new techniques to visualize and animate requirements models [...] beautifully simple but potentially very effective" [20]. Such folk knowledge assumes that a graphical RE model would be anyhow better. This seemed to be partly confirmed by our initial experiment in 2013 [10]. Yet, our other experiments failed to produce strong, conclusive evidence in this respect [9,8].

The literature suggests that tabular methods support better the identification of threats and security requirements than graphical methods. Opdahl and Sindre [21] compared misuse cases with attack trees in a controlled experiment with students and repeated it with industrial practitioners in [22]. Both studies showed that attack trees help to identify more threats than misuse cases, but both methods have similar perception. Stålhane et al. have conducted a series of experiments to evaluate two representations of misuse cases: a graphical diagram and a textual template. The results reported in [23] revealed that textual use cases helped to identify more threats than use-case diagrams. In more recent experiments [24,25,26], Stålhane et al. compared textual misuse cases with UML system sequence diagrams. The results showed that textual misuse cases are better than sequence diagrams in identification of threats related to required functionality or user behavior. In contrast, sequence diagrams outperform textual use cases in the identification of threats related to the system's internal working. Scandariato et al. [27] evaluated Microsoft STRIDE [5], which is is

a mix of graphical (Data Flow Diagrams) and tabular notations. The results showed that STRIDE is not perceived as difficult by the participants but their productivity in threats identified per hour was very low. Besides, the correctness of the threat is good because the participants identified only few incorrect threats but the completeness was low because they overlook many threats.

3 Research Design

To answer our research questions we cannot use the approach used in the previous papers [10,9,8] as they attempted to prove difference and the lack of evidence for difference is not the same as evidence for equivalence. Hence, we use **equivalence testing** – TOST, which was initially proposed by Schuirmann [12] and is widely used in pharmacological and food sciences to answer the question whether two treatments are equivalent within a particular range δ [28,13]. We summarize the key aspects of TOST as it is not well known in SE and refer to [13] for details. The problem of the equivalence test can be formulated as follows:

$$H_0 : |\mu_A - \mu_B| > \delta \quad \text{vs} \quad H_a : |\mu_A - \mu_B| \leq \delta. \tag{1}$$

where μ_A and μ_B are means of methods A and B, and δ corresponds to the range within which we consider two methods to be equivalent.

Such question can be tested as a combination of *two* tests, as:

$$\begin{array}{l} H_{01} : \mu_A < \mu_B - \delta \quad \text{or} \quad H_{02} : \mu_A > \mu_B + \delta \\ H_{a1} : \mu_A \geq \mu_B - \delta \quad \text{and} \quad H_{a2} : \mu_A \leq \mu_B + \delta, \end{array} \tag{2}$$

The p-value is then the maximum among p-values of the two tests (see [13] for an explanation on why it is not necessary to perform a Bonferroni-Holms correction). The underlaying statistical test for each of these two alternative hypothesis can then be any difference tests (eg. t-test, Wilcoxon, Mann-Whitney etc.) as appropriate to the underlying data.

For variables collected along a 1-5 Likert scale, a percentage test [28] may grant statistical equivalence too easily and, therefore, we ran an absolute test with narrower range of $\delta = \pm 0.6$. A statistical difference would then correspond to a clear practical difference: a gap in the perception of two methods bigger than > 0.6 means that around $2/3$ of participants ranked one method at least one point higher than the rank of the other method. For the qualitative evaluation of the security assessment by the experts it means that, e.g., two out of three experts gave one point higher to SRA performed with one method comparing to the results of the other method. It corresponds to 20% range on a 5-item scale with mean value equal to 3.

Study Design and Planning. We chose a *within-subject design* where each group applied both methods. To avoid limitations due to domain security knowledge, each group was also given a professional-level domain-specific catalogue (its effects are described in [29]). To avoid learning effects, each group was asked to perform the risk assessment for a different security task in the same domain. Table 1 summarizes the *treatment variables* that we used in our study.

In our study each group performed the risk analysis of four security tasks (see Table 1). To control the effect of security tasks on results we split groups

Table 1: Experimental Variables

As treatments we had two methods, four security tasks, and two experiments. As dependent variables we had quality of threats and security control as a measure of actual efficacy, and PEOU and PU as a measure of method's perception.

Type	Name	Description
Treatment	Tabular, Graphical	The method used to conduct SRA for a security task: SESAR SecRAM (Tabular) or CORAS (Graphical).
	IM, AM, WebApp/DB, and Network	The groups have to conduct SRA for each of four security tasks: *1)* Identity Management (IM) and *2)* Access Management (AM) Security, *3)* Web Application and Database Security (WebApp/DB), and *1)* Network and Infrastructural Security (Network).
	Experiment X	The study consisted of two controlled experiments: Experiment 1 and Experiment 2.
Actual Efficacy	Q_T, Q_{SC}	The overall quality of threats (Q_T) and security controls (Q_{SC}) based on the evaluation from three independent security experts.
Perceived Efficacy	PEOU, PU	Mean of the responses to the eight questions about perceived ease of use (PEOU) and nine questions about perceived usefulness (PU).

into two types: type A groups started by using the graphical method on IM, then the tabular method on AM and so on, alternating methods, while type B groups did the opposite. Each group was randomly assigned to either type A or B.

Experimental Protocol. Our protocol consists of three main phases:

Training. Participants were administered a short demographics and background questionnaire. For each SRA method and application scenario participants attended 2h lecture given by an author of the paper. Each lecture on method was followed by a practical exercise on a toy scenario demonstrating application of the corresponding method. After, participants were divided in groups of two and received training materials including EUROCONTROL EATM security catalogues and scenario description. Since catalogues and ROT description are confidential materials for EUROCONTROL, participants received only a paper version of the documents and had to sign a non-disclosure agreement.

Application. Once trained on the scenario and methods, groups had to apply each method to four different tasks (two per method). For each task, groups:

- Attended a two hours lecture on the threats and possible security controls specific to the task but not specific to the scenario.
- Had 2 weeks to apply the assigned methods to identify threats and security controls specific for the task.
- Delivered an intermediate report.
- Gave a short presentation about the preliminary results of the method application and received feedback from one of the authors of this paper.

Evaluation. Three experts independently evaluated the quality of threats and security controls identified by groups and the overall quality of the report, providing marks and justifications. Participants received experts' assessments and the course final mark. After, they were asked to answer the post-task questionnaire to collect their perception of the methods taking into account the feedback.

Data Collection. Table 1 reports *dependent variables* for actual and perceived efficacy. To answer *RQ1* we measured a method's *actual efficacy* by ask-

ing external security experts to independently evaluate the quality of identified threats and security controls for each security task on a five-item scale: *Bad* (1), *Poor* (2), *Fair* (3), *Good* (4), and *Excellent* (5). Such choice is motivated by several factors. At first, the quality of results is considered to be more important in practice: "the security risk assessment report is expected to contain adequate and *relevant* evidence to support its findings, clear and *relevant* recommendations" [30] (Our emphasis). Second, as all participants were provided with a catalogues, they could easily produce a large number of threats and control, irrespective of the method used. Further, [21] have also reported that different methods might help to generate outcomes of difference quality: participants using attack trees identified mainly generic threats, while misuse cases helped to identify more domain-specific threats.

To answer *RQ2* we collected participants' opinion PEOU and PU of both methods using a post-task questionnaire at the very end of our study. The post-task questionnaire was inspired by the Technology Acceptance Model (TAM) [31] and a similar questionnaire used in [21,9]. The questions were formulated in one sentence with answers on a 5-point Likert scale (1 - Strongly agree; 2 - Agree; 3 - Not certain; 4 - Agree; 5 - Strongly agree)[3].

In [10,9] we used raw responses to individual questions within each category to compare PEOU and PU of two methods. Karpati et al. [22] used the mean of participants' responses to PEOU and PU questions as a consolidated measure of their PEOU and PU. The approach by Karpati et al. seems to be more robust against the possible fluctuation of the responses within the same category. Therefore, in the current study we adopted this approach.

Data Analysis. To test for statistical difference, we used the following underlying non-parametric tests for difference as our data is ordinal and not normal:

- Mann-Whitney (MW) test to compare two unpaired groups (eg. quality of threats in two experiments).
- Wilcoxon signed-rank test to compare two paired groups (eg. participants' perception of two methods).
- Kruskal-Wallis (KW) test to compare more than two unpaired groups (eg. quality of threats in four security tasks).
- Spearman's rho coefficient for correlation.

For the hypotheses about equivalence of two treatments we applied TOST with Wilcoxon test as the underlying test. The TOST and selection of the equivalence range is discussed in Section 3. For all statistical test we adopted 5% as a threshold for α (i.e. probability of committing Type-I error) [32].

4 Study Realization

The study consisted of two controlled experiments: Experiment 1 and 2. The participants of the study were MSc students enrolled to Security Engineering course taught by one of the author in Fall semesters of 2014-2015 and 2015-2016

[3] To prevent participants from "auto-pilot" answering, a half of the questions were given in a positive statement and another half in a negative statement.

Table 2: Overall participants' Demographic Statistics

Experiment 1

Variable	**Scale**	**Mean/ Median**	**Distribution**
Age	Years	23.1	43.3% were 19-22 years old; 43.3% were 23-25 years old; 13.3% were 26-31 years old
Gender	Sex		75.8% male; 24.2% female
Work Experience	—	1.3	46.7% had no experience; 36.7% had 1-2 years; 13.3% had 3-5 years; 3.3% had 6 years
Expertise in Security	0(Novice)-4(Expert)	1 (median)	26.7% novices; 60% beginners; 13.3% competent users
Expertise in Modeling Languages	—	1 (median)	26.7% novices; 26.7% beginners; 40% competent users; 6.7% proficient users
Expertise in ATM	—	0 (median)	93.3% novices; 6.7% beginners

Experiment 2

Variable	**Scale**	**Mean/ Median**	**Distribution**
Age	Years	24.4	32.6% were 21-22 years old; 34.9% were 23-25 years old; 32.6% were 26-30 years old
Gender	Sex		78.3% male; 21.7% female
Work Experience	—	2.1	23.3% had no experience; 44.2% had 1-2 years; 23.3% had 3-5 years; 9.3% had 6-10 years
Expertise in Security	0(Novice)-4(Expert)	1 (median)	30.2% novices; 41.9% beginners; 11.6% competent users; 11.6% proficient users; 4.7% experts
Expertise in Modeling Languages	—	1 (median)	11.6% novices; 41.9% beginners; 30.2% competent users; 16.3% proficient users
Expertise in ATM	—	0 (median)	69.8% novices; 27.9% beginners; 2.3% competent users

academic years at the University of Trento, Italy. Experiments involved 35 and 48 participants correspondingly. Participants worked in groups of 2 members, except one participant in Experiment 1 who did not have a partner. We had to discard the results from 5 participants in Experiment 1 and 2 participants in Experiment 2 because they failed to complete all necessary steps of the study or provide inconsistent responses to a post-task questionnaire. If the problem was only with post-task questionnaire, we discarded the results only from $RQ2$ analysis and kept the group's results in the analysis for $RQ1$.

Table 2 reports participants' demographics in Experiment 1 (above) and 2 (below). A half of the participants (53.3%) in Experiment 1 and most participants (76.7%) in Experiment 2 reported that they had working experience. In Experiment 1 the participants had basic knowledge of security, while in Experiment 2 the participants reported good general knowledge of security. In both experiments the participants had basic knowledge of modeling languages and limited background in the application scenario.

Application Scenario Selection. In Experiment 1 as an application scenario we selected the Remotely Operated Tower (ROT) which was developed for and used in our previous study [29]. ROT is a new operational concept proposed by SESAR in order to optimize the air traffic management in the small and remote airports. The main idea is that control tower operators will no longer be located at the airport. The air traffic controllers will use a graphical reproduc-

tion of the out-of-the-window view by means of cameras with a 360-degree view which overlaid with information from other sources like surface movement radar, surveillance radar, and others. The first implementation of ROT has been done by LFV and Saab in Sweden in 2015 [4].

To control the possible "learning effects" between different experiments, in Experiment 2 we switched to the application scenario on the Unmanned Aerial System Traffic Management (UTM) based on the documents from NASA [33], Amazon's memorandum for commercial interests [34], and the thesis on the integration of drones into the national aerospace system [35].

Tasks. For both application scenarios we asked our groups to conduct SRA for each security task (see Table 1) using the corresponding method according to the predefined order. For example, in WebApp/DB task they could identify threats like SQL injection or DoS attack and propose controls to mitigate them.

Methods Selection. In this study we continued our work reported in [10,9]. Thus, as an instance of graphical method we kept CORAS method *a)* in order to have a common point of comparison with the previous studies and *b)* because it provides a clear process to conduct SRA. CORAS was design by SINTEF [3], a research institution in Norway. They use this method to provide consulting services to their clients. CORAS is a *graphical method* whose analysis is supported by a set of diagrams that represent assets, threats, risks and treatments. This method supports both the ISO 27005 and ISO 31000 standards.

In contrast to [9], as a tabular methods we selected another ATM Security Risk Assessment Method (SecRAM) proposed by SESAR. The method was developed within project 16.02.03 [11], and used by professionals in the SESAR program to conduct SRA. This method was designed as an easy to use step-wise method that can be applied to any operational focus ares of SESAR. Which means that its process should be clearer comparing to EUROCONTROL SecRAM's process. Further when we use SecRAM we refer to SESAR SecRAM unless otherwise stated.

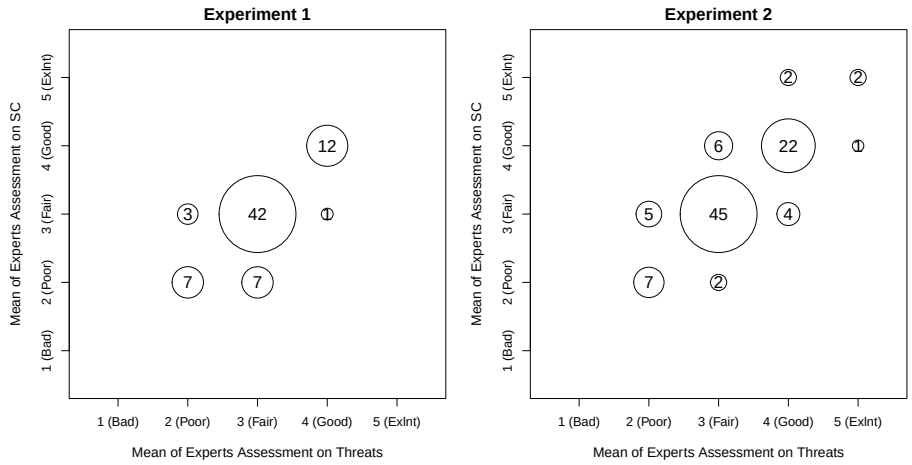
5 Results

In the following, we report results of our study, with the aim of answering the research questions formulated in Section 3.

First, we performed an analysis on the various experimental factors (i.e. experiments and tasks) to determine whether there was a significant difference. Factors without a significant difference in outcomes were aggregated, whereas outcomes for factors with a significant difference were reported separately.

Factor - Security Task: The results of pairwise TOST with Wilcoxon test confirmed the equivalence of each pair of tasks for the quality of threats (p-$value < 0.021$ in Exp. 1 and p-$value < 0.002$ in Exp. 2) and controls (p-$value < 0.004$ in Exp. 1 and p-$value < 2*10^{-5}$ in Exp. 2). Therefore, we can use the mean quality of threats and controls identified for two tasks as a measure of actual

[4] LFV: RTS - One Year In Operation. Available: `http://news.cision.com/lfv/r/rts---one-year-in-operation,c9930962`



The figures report experts overall quality assessment of the threats and controls identified for four security tasks in Experiment 1 (left) and 2 (right). The majority of the groups delivered threats and controls of "fair" and "good" quality. Only limited number of the reports delivered "poor" threats and security controls. The quality of the results was better than previous experiments [9] and we did not split groups into "good" and "bad".

Fig. 1: Experts assessment by methods and experiments

efficacy for a method. In this way we can eliminate a possible effect of task order on the results of Wilcoxon test and compare paired data.

Factor - Experiment: The results of TOST confirmed the equivalence of two experiments for the mean quality of threats and controls for both methods (TOST *p-value* < 0.005). However, TOST failed to reject the hypothesis about non-equivalence of two experiments for the mean participants' PEOU (TOST *p-value* $= 0.21$) and PU (TOST *p-value* $= 0.07$) for graphical method. Hence, we report the results of the two experiments separately.

Factor - Background: In both experiments the KW test did not revealed any statistically significant effect of background variables (see Table 2) on the quality of threats and controls or mean participants' PEOU and PU.

RQ1: Actual Efficacy. Figure 1 reports the mean of experts assessment of threats and security controls identified by groups. In Experiment 1 and 2 we had 18 and respectively 24 groups that successfully delivered the final report and were evaluated by the experts. In total we collected 72 methods applications in Experiment 1 and 96 in Experiment 2. The overall quality of the identified threats and security controls was "fair" or "good".

Table 3 presents the descriptive statistics, p-values of the TOST with Wilcoxon test for the equivalence in the mean quality of threats and controls by experiment and method. In Experiment 1 tabular method helped to identify threats and controls of a slightly better quality than the graphical one and in Experiment 2 both methods helped to produce results of the same quality. However, in both

Table 3: Average quality of threats and sec. controls by experiments and methods

Tabular and graphical methods produces very similar quality of threats and controls in both experiments. The quality of the produced threat is within a 10% range around the mean quality range (3 - fair). For both experiments this is statistically significant with a TOST for an effect size of $\delta = \pm 0.6$ corresponding to less than two experts having a different rate of the outcome of the risk assessment.

	Actual	Tabular			Graphical			δ_{mean}	TOST
	Efficacy	Mean	Median	St. dev.	Mean	Median	St. dev.	Tab - Graph	p-value
Exp. 1	Threats	3.17	3.08	0.53	2.95	2.92	0.53	+0.22	0.0009
	Sec. Ctrls	3.28	3.25	0.53	2.97	2.92	0.51	+0.31	0.001
Exp. 2	Threats	3.28	3.17	0.58	3.24	3.17	0.57	+0.04	$6.3 \cdot 10^{-6}$
	Sec. Ctrls	3.31	3.25	0.67	3.29	3.25	0.62	+0.02	$2.4 \cdot 10^{-7}$

Table 4: Average perception of tabular and graphical SRA methods

The results of Experiment 1 showed that the participants reported higher PEOU and PU for the tabular method than for the graphical one. However, TOST ($\delta = \pm 0.6$) results did not reveal any equivalence of two methods and Wilcoxon results did not confirm that the difference is statistically significant. The results of Experiment 2 revealed that two methods are equivalent with respect to PEOU (statistically significant with a TOST for an effect size of $\delta = \pm 0.6$).

	Perceived	Tabular			Graphical			δ_{mean}	TOST
	Efficacy	Mean	Median	St. dev.	Mean	Median	St. dev.	Tab - Graph	p-value
Exp.1	PEOU	3.63	3.75	0.59	3.20	3.12	0.64	+0.43	0.08
	PU	3.54	3.72	0.84	3.05	3.17	0.83	+0.37	0.18
Exp. 2	PEOU	3.74	3.75	0.40	3.60	3.69	0.71	+0.14	$2.6 \cdot 10^{-5}$
	PU	3.67	3.78	0.58	3.29	3.44	0.99	+0.38	0.03

experiments the TOST results confirmed that the *two methods are equivalent in threats and controls quality.*

RQ2: Perceived Efficacy. Table 4 reports the descriptive statistics, p-values of TOST with Wilcoxon test for the equivalence in participants' PEOU and PU by experiment and method. In Experiment 1 the participants reported better perception of the tabular model over the graphical one for all three variables. Such difference in mean was lower than our TOST practical significance threshold of $\delta = \pm 0.6$. TOST failed to reject the hypotheses about non-equivalence between two methods for PEOU and PU. In Experiment 2 we have different picture. The perception of the graphical method significantly increased in this experiment comparing to the first one. So, the two methods have equivalent PEOU and PU which confirmed by TOST results.

6 Retrospective Analysis

In the previous studies [10,9] we compared graphical method CORAS with different tabular methods. In [10] as a tabular method we chose SREP [6] proposed by University of Castilla–La Mancha and used by CMU Software Engineering Institute in their tutorials. The participants worked in groups of two and conducted SRA of four security tasks from SmartGrid scenario using both methods. The division of groups on good and "not good" was done based on security experts assessment of the final reports quality. In [9] we used tabular method from in-

dustry proposed by EUROCONTROL, SecRAM. The participants individually conducted SRA of two tasks from SmartGrid scenario using both methods.

In [10,9] we followed the approach by Opdahl and Karpati [21] and used the number of threats and security controls identified using a method as a measure of the actual efficacy. Thus, we cannot compare current results with the results reported in [10,9], but this comparison can be done for the perception variables.

We re-ran hypothesis testing for the equivalence of two methods in participants' PEOU and PU using TOST with MW test. We chose MW test to have comparable results across all experiments as we cannot used Wilcoxon test when we analyze the results of good groups where the samples can be unpaired.

The results of the retrospective analysis do not contradict to the results reported in the first study [10]. For good groups TOST failed to reject the hypothesis about non-equivalence in mean PEOU (p-$value$= 0.25) and PU (p-$value$= 0.27). Across all groups TOST results confirmed the equivalence of two methods w.r.t. mean PEOU (p-$value$ = 0.051) and PU (p-$value$ = 0.003).

For the second study [9] the retrospective analysis for all participants revealed: *a)* 10% significantly better mean PEOU in favor of graphical method (MW p-$value$ = 0.06) and *b)* 10% significant equivalence in mean PU between two methods (TOST p-$value$ = 0.08). For good participants TOST failed to reject hypothesis about non-equivalence of two methods in mean PEOU (p-$value$ = 0.85) and PU (p-$value$ = 0.43). Slightly better PEOU of the graphical method might be because its process is clearer comparing to the process of the tabular method as suggested by the qualitative analysis results (see Table III in [9]).

The difference between the results reported for the perception in [9] and the results of the retrospective analysis can be due to the different data collection approach which is discussed in Section 3.

7 Discussion

The results showed in both experiments that two methods are equally good in terms of quality of identified threats and controls. Thus, in response to $RQ1$ we can conclude that *tabular and graphical methods are equal w.r.t. actual efficacy.*

In Experiment 1 we observed slightly better participants' PEOU and PU, but the results failed to reveal any statistically significant equivalence or difference between two methods in these variables. At the same time, in Experiment 2 tabular and graphical methods were found to be statistically equivalent in terms of participants' PEOU and PU. The possible explanation is that in Experiment 1 we used the ROT scenario that was designed by the same organization which designed tabular method and security catalogues. Possible this combination is a "good fit" which led to better perception of the tabular method. In Experiment 2 we used UATM scenario by NASA that might be "not a good fit" to the same combination of tabular method and security catalogues. This could result in a similar perceived ease of use and usefulness.

Another possible factor that impacted participants' PEOU and led to its significant improvement are the changes in the feedback process between two

experiments. In Experiment 1 the public discussion of groups' deliverables was *at will* and it might happened that not all groups decided to use their possibility to discuss the work. Besides the discussion in the class, each group received *individual feedback* on the mistakes of method application found in their deliverables. In contrast, in Experiment 2 we allocated 15 min slots and asked groups to register for the open feedback session in advance. Each group participated in *at least one feedback session* and gave a 5 min presentation on the intermediate results. Besides the discussion by groups, for each deliverable we provided groups with *the summary of the typical problems* in the application of both methods.

Moreover, in Experiment 2 we provided *feedback on typical mistakes* that the participants did in the *warm-up SRA* of a toy application scenario that was a part of the training. So, the groups were able to better understand the methods and avoid mistakes from the very first deliverable. In Experiment 1 such feedback on the warm-up exercise was not provided. We could expect that these changes would also result in a significant increase of tabular method's PEOU, but this method has simpler process and representation requiring less effort to master comparing to the graphical method. Therefore, tabular method's PEOU did not significantly increase in the second experiment.

The results of retrospective analysis of the previous experiments supports the current findings. In [10] graphical and tabular methods have similar PEOU and PU as graphical and tabular methods have clear process. In contrast, in [9] the graphical method has higher PEOU than the tabular one because graphical method has significantly clearer process comparing to the tabular method.

The answer to $RQ2$ is: *if there is no fit between SRA components (i.e. method, catalogues, and application scenario) and methods have equally clear processes then there will be no difference in perceived efficacy of these methods.* However, if a method *a)* operates same concepts and terminology as an application scenario, and/or *b)* has very clear process, then it may result in a better perceived efficacy for this method comparing to the other methods.

8 Threats to Validity

In this section we discuss the main types of threats to validity [32].

Regarding **internal validity**, the main concern is that the relations between the treatment and the outcome are causal and the effects of possible factors are either controlled or measured. To mitigate this we randomly assigned groups to the order of methods application. The results of two experiments were reported and discussed separately to alleviate the possible effect of the differences in experiments execution. The results of KW test did not reveal any statistically significant effect of participants' background and experience on the results.

The main threats to **construct validity** are the definition and interpretation of the metrics that we used to measure the theoretical constructs. We measured the *actual efficacy* of a method as the quality of threats and security controls identified using a method. The relevance of results quality for an SRA is discussing in Section 3. To measure the *perceived efficacy* we designed the post-task

questionnaires following TAM [31]. The questionnaire includes 8 questions about PEOU and 9 questions about PU, which were adapted from [10,9].

A main threat to **conclusion validity** is related to *low statistical significance* of the findings. The effect size for the equivalence test was set to $\delta = \pm 0.6$ which corresponds to 20% difference in actual or perceived efficacy. The practical meaning of this threshold is discussed in Section 3.

In regard to **external validity**, the main issues threatening the generalizability of the results are the *use of students instead of practitioners* and the use of *simple scenarios* to apply the methods under evaluation [36]. The use of MSc students in empirical studies is still question of debate. However, some studies have argued that students perform as well as professionals [37,38]. Regarding the use of simple scenarios, in our studies we mitigated this threat by asking the participants to analyze two new operational scenarios introduced in the ATM domain.

9 Conclusion

In our previous studies [10,9] with a similar settings, i.e. full-scale application of tabular and graphical methods, the results did not reveal consistent superiority of one method in identification of threats and/or controls. However, the graphical method was reported to have higher perception than the tabular one. The possible explanation could be that we were looking for the difference between two methods without defining *how big the difference* should be in order to *be different*, which is not envisaged by the hypothesis tests like Wilcoxon and Mann-Whitney.

Instead, in this study we decided to investigate how similar are security methods with respect to actual and perceived efficacy. The difference range we defined in terms of $\delta = \pm 20\%$. For example, for 5-item Likert scale for quality/perception value the δ is equal to ± 0.6. It means, for example, that tabular and graphical methods are equivalent in terms of threats quality if $|Q(T_{graph}) - Q(T_{tab}| < \delta$.

The results of the two controlled experiments revealed that tabular and graphical methods are equivalent in terms of *actual efficacy* ($RQ1$). The groups were able to identify threats and controls of a fair quality with both methods.

Regarding the difference in *methods' perception* ($RQ2$), the data analysis results showed that participants perceived tabular method to be slightly better with respect to *perceived ease of use and usefulness* than the graphical one in the first experiments, and in the second experiment the two methods were found to be statistically equivalent with respect to perception variables.

To summarize, the study shows that tabular and graphical methods for (security) requirements elicitation and risk assessment are very similar with respect to actual and perceived efficacy. Graphical representation only does not guarantee the better quality of security requirements analysis in comparison to a tabular method.

References

1. A. Van Lamsweerde, "Goal-oriented requirements engineering: A guided tour," in *Proc. of RE 2001*. IEEE, 2001, pp. 249–262.
2. P. Giorgini, F. Massacci, J. Mylopoulos, and N. Zannone, "Modeling security requirements through ownership, permission and delegation," in *Proc. of RE 2005*. IEEE, 2005, pp. 167–176.
3. M. S. Lund, B. Solhaug, and K. Stølen, "A guided tour of the CORAS method," in *Model-Driven Risk Analysis*. Springer, 2011, pp. 23–43.
4. R. Caralli, J. Stevens, L. Young, and W. Wilson, "Introducing OCTAVE allegro: Improving the information security risk assessment process," Software Engineering Institute, Carnegie Mellon University, Tech. Rep., 2007.
5. 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.
6. D. Mellado, E. Fernández-Medina, and M. Piattini, "Applying a security requirements engineering process," in *Proc. of ESORICS'06*. Springer, 2006, pp. 192–206.
7. F. Massacci and F. Paci, "How to select a security requirements method? a comparative study with students and practitioners," in *Proc. of NordSec 2012*. Springer, 2012, pp. 89–104.
8. K. Labunets, F. Paci, F. Massacci, M. Ragosta, and B. Solhaug, "A First Empirical Evaluation Framework for Security Risk Assessment Methods in the ATM Domain," in *Proc. of SIDs 2014*. SESAR, 2014.
9. K. Labunets, F. Paci, F. Massacci, and R. Ruprai, "An experiment on comparing textual vs. visual industrial methods for security risk assessment," in *Proc. of EmpiRE Workshop at RE 2014*. IEEE, 2014, pp. 28–35.
10. K. Labunets, F. Massacci, F. Paci, and L. M. S. Tran, "An Experimental Comparison of Two Risk-Based Security Methods," in *Proc. of ESEM 2013*. IEEE, 2013, pp. 163–172.
11. SESAR, *ATM Security Risk Assessment Methodology. SESAR WP16.2 ATM Security*, February 2003.
12. D. Schuirmann, "On hypothesis-testing to determine if the mean of a normal-distribution is contained in a known interval," in *Biometrics*, vol. 37, no. 3. International Biometric Soc, 1981, pp. 617–617.
13. M. Meyners, "Equivalence tests–a review," *Food quality and preference*, vol. 26, no. 2, pp. 231–245, 2012.
14. E. S. Yu, "Modeling organizations for information systems requirements engineering," in *Proc. of RE 1993*. IEEE, 1993, pp. 34–41.
15. S. Robertson and J. Robertson, *Mastering the requirements process: Getting requirements right*. Addison-wesley, 2012.
16. T. Li and J. Horkoff, "Dealing with security requirements for socio-technical systems: A holistic approach," in *Proc. of CAiSE'14*. Springer, 2014, pp. 285–300.
17. H. Mouratidis and P. Giorgini, "Secure tropos: a security-oriented extension of the tropos methodology," *International Journal of Software Engineering and Knowledge Engineering*, vol. 17, no. 02, pp. 285–309, 2007.
18. C. Haley, R. Laney, J. Moffett, and B. Nuseibeh, "Security requirements engineering: A framework for representation and analysis," *IEEE T. Software Eng.*, vol. 34, no. 1, pp. 133–153, 2008.
19. M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen, "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements," *Req. Eng.*, vol. 16, no. 1, pp. 3–32, 2011.

20. N. Maiden, S. Robertson, and C. Ebert, "Guest editors' introduction: Shake, rattle, and requirements," *IEEE Software*, vol. 22, no. 1, p. 13, 2005.
21. A. L. Opdahl and G. Sindre, "Experimental comparison of attack trees and misuse cases for security threat identification," *Inform. Soft. Tech.*, vol. 51, no. 5, pp. 916–932, 2009.
22. P. Karpati, Y. Redda, A. L. Opdahl, and G. Sindre, "Comparing attack trees and misuse cases in an industrial setting," *Inform. Soft. Tech.*, vol. 56, no. 3, pp. 294–308, 2014.
23. T. Stålhane and G. Sindre, "Safety hazard identification by misuse cases: Experimental comparison of text and diagrams," in *Proc. of MODELS 2008*, 2008, pp. 721–735.
24. T. Stålhane, G. Sindre, and L. Bousquet, "Comparing safety analysis based on sequence diagrams and textual use cases," in *Proc. of CAiSE'10*, vol. 6051, 2010, pp. 165–179.
25. T. Stålhane and G. Sindre, "Identifying safety hazards: An experimental comparison of system diagrams and textual use cases," in *Proc. of BPMDS 2012*, vol. 113, 2012, pp. 378–392.
26. T. Stålhane and G. Sindre, "An experimental comparison of system diagrams and textual use cases for the identification of safety hazards," *International Journal of Information System Modeling and Design*, vol. 5, no. 1, pp. 1–24, 2014.
27. R. Scandariato, K. Wuyts, and W. Joosen, "A descriptive study of Microsoft's threat modeling technique," *Req. Eng.*, pp. 1–18, 2014.
28. Food and Drug Administration, "Guidance for industry: Statistical approaches to establishing bioequivalence," 2001.
29. 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 2015*, ser. Lecture Notes in Computer Science, vol. 9013. Springer, 2015, pp. 98–114.
30. D. J. Landoll and D. Landoll, *The security risk assessment handbook: A complete guide for performing security risk assessments*. CRC Press, 2005.
31. F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quart.*, pp. 319–340, 1989.
32. C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Springer, 2012.
33. P. H. Kopardekar, "Unmanned aerial system (UAS) traffic management (UTM): Enabling low-altitude airspace and UAS operations," Tech. Rep., 2014.
34. ——, "Revising the airspace model for the safe integration of small unmanned aircraft systems," Tech. Rep., 2015.
35. C. A. Theilmann, "Integrating autonomous drones into the national aerospace system," Ph.D. dissertation, University of Pennsylvania, PA, US, April 2015.
36. J. C. Carver, L. Jaccheri, S. Morasca, and F. Shull, "A checklist for integrating student empirical studies with research and teaching goals," *Empirical Software Engineering*, vol. 15, no. 1, pp. 35–59, 2010.
37. M. Svahnberg, A. Aurum, and C. Wohlin, "Using students as subjects - an empirical evaluation," in *Proc. of ESEM 2008*. ACM, 2008, pp. 288–290.
38. M. Höst, B. Regnell, and C. Wohlin, "Using students as subjects: A comparative study of students and professionals in lead-time impact assessment," *Empirical Softw. Engg.*, vol. 5, no. 3, pp. 201–214, Nov. 2000.