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 [35] to [6] and more recently [19]. Industry opts for tabular models like OCTAVE [1], ISO 27005 and NIST 800-30. Microsoft STRIDE [9] is the exception on the industry side and SREP [22] is the exception on the academic side.

The initial goal of our long term experimental plan in 2011 [21] 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 [15,16], 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 learning effect 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[3] 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 SG2013 study) 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 [22] which we used in SG2013 study by a method used in the industry SecRAM [28] which had very similar tables but a nimbler process, designed by risk-assessment industry experts to simplify SRA, in the same fashion that the graphical method was designed by SINTEF to be simple to use in its industry consultancies [19].

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) [27,23] 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 while the presence of clear process may improve method's perception.

[3] For simplicity, we name our previous experiments as "SG2013" [14] and "SG2014" [16], where *SG* stands for Smart Grid domain used in the experiments.

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 [35,6], others offered variants of graphical notation [18,24,8,19], or minimal model based transformation analysis [4]. 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 [6].

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 [28], 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 [21] 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 [6] 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 "SG2013". Yet, our other experiments failed to produce strong, conclusive evidence in this respect [16,15].

Empirical Comparison of Graphical and Tabular Representations

Graphical and textual notations were empirically investigated in different domains. In this discussion we focused on the works similar to ours that investigate these representations in security requirements engineering.

Opdahl and Sindre [25] compared misuse cases with attack trees in a controlled experiment with students and repeated it with industrial practitioners in [11]. Both studies used Wilcoxon signed-ranks test for difference between two methods. The results 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. In these experiment authors used t-tests to compare two representations. The results reported in [29] revealed that textual use cases helped to identify more threats than use-case diagrams. In more recent experiments [32,30,31], 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. [26] evaluated Microsoft STRIDE [9], which is is a mix of graphical (Data Flow Diagrams) and tabular notations. The authors used Wilcoxon test to compare different aspects of the methodology. 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.

To answer our research questions we cannot use the standard statistical tests (e.g. t-test, Wilcoxon, etc.) as they attempt to prove difference and the lack of evidence for difference is not the same as evidence for equivalence.

3 Research Design

We use **equivalence testing** – TOST, which was proposed by Schuirmann [27] and is widely used in pharmacological and food sciences to answer the question whether two treatments are equivalent within a particular range δ [5,23]. We summarize the key aspects of TOST as it is not well known in SE and refer to the review paper by Meyners [23] 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 [23] 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 [5] 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.

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: ROT2015 and UTM2016.
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).

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. We showed that catalogues are effective in equalizing non-security experts and security experts (without a catalogue) in [7]. To avoid learning effects, each group was asked to perform SRA for a different security task in the same domain. Table 1 summarizes *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 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. Next, 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. Finally, 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 asking 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" [17] (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, [25] 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) [3] and a similar questionnaire used in [25,16]. 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)[4]. We followed the approach by Karpati et al. [11] and used the mean of participants' responses to PEOU and PU questions as a consolidated measure of their PEOU and PU. This approach seems to be more robust against the possible fluctuation of the responses within the same category.

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) [36].

[4] 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 ROT2015

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 UTM2016

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

4 Study Realization

The study consisted of two controlled experiments: ROT2015 and UTM2016. 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 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 ROT2015 who did not have a partner. We had to discard the results from 5 participants in ROT2015 and 2 participants in UTM2016 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 ROT2015 (above) and UTM2016 (below). A half of the participants (53.3%) in ROT2015 and most participants (76.7%) in UTM2016 reported that they had working experience. In ROT2015 the participants had basic knowledge of security, while in UTM2016 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 ROT2015 as an application scenario we selected the Remotely Operated Tower (ROT) which was developed for and

used in our previous study [7]. 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 reproduction 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 [5].

To control the possible "learning effects" between different experiments, in UTM2016 we switched to the application scenario on the Unmanned Aerial System Traffic Management (UTM) based on the documents from NASA [12], Amazon's memorandum for commercial interests [13], and the thesis on the integration of drones into the national aerospace system [34].

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 [14,16]. 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 [19], 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 and provides guidance through 8-steps SRA process: *1)* preparation for the analysis, *2)* customer presentation of the target, *3)* approval of the target description, *4)* refining the target description, *5)* risk identification, *6)* risk estimation, *7)* risk evaluation, and *8)* risk treatment.

As a tabular methods we selected another ATM Security Risk Assessment Method (SecRAM) developed by SESAR (Single European Sky ATM Research Program) within 16.02.03 project[6]. The method was 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. Further when we use SecRAM we refer to SESAR SecRAM unless otherwise stated. SecRAM process includes 7 main steps: *1)* primary assets identification and impact assessment, *2)* supporting assets identification and evaluation, *3)* vulnerabilities and threats identification, *4)* likelihood evaluation, *5)* impact evaluation, *6)* risk level evaluation, and *7)* risk treatment. SecRAM uses tables to represent results of each step.

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

[6] SESAR Project 16.02.03 - ATM Security Risk Assessment Methodology, February 2003. Project aims to analyze existing security risk assessment approaches and adopt them to the ATM domain.

5 Results

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 ($p< 0.021$ in ROT2015 and $p< 0.002$ in UTM2016) and controls ($p< 0.004$ in ROT2015 and $p< 2*10^{-5}$ in UTM2016). Therefore, we can use the mean quality of threats and controls identified for two tasks as a measure of actual 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 $p < 0.005$). However, TOST failed to reject the hypothesis about non-equivalence of two experiments for the mean participants' PEOU (TOST $p = 0.21$) and PU (TOST $p = 0.07$) for graphical method. Hence, we report the results of the two experiments separately.

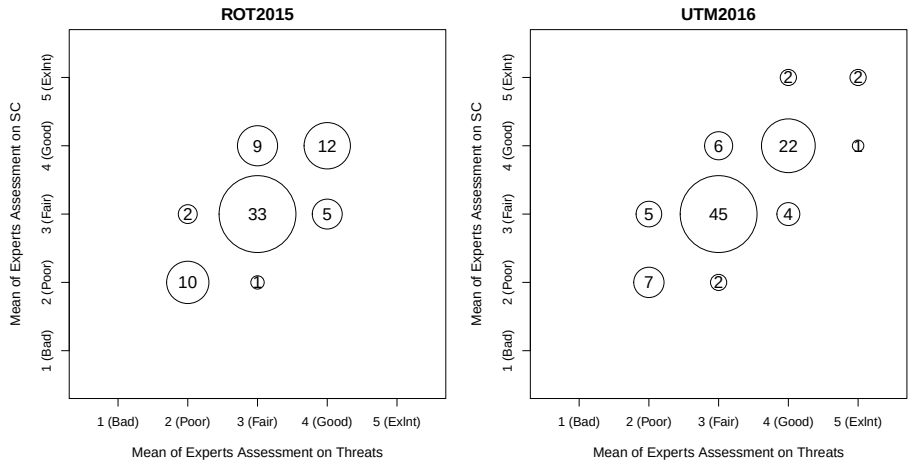
Factor - Background: In both experiments the KW test did not revealed any statistically significant effect of background variables (see Table 2) on the quality of threats and controls or mean participants' PEOU and PU.

RQ1: Actual Efficacy. Figure 1 reports the mean of experts assessment of threats and security controls identified by groups. In ROT2015 and UTM2016 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 ROT2015 and 96 in UTM2016. The overall quality of the identified threats and security controls was "fair" or "good".

To provide an idea on the scale of results produced by groups, we report the number of threats and security controls identified by one of the best groups for each experiment. In ROT2015 the best group identified in total 49 threats and 120 security controls which composed 178 pairs as some controls can be used to mitigate different threats. In UTM2016 the best groups identified totally 53 threats and 36 security controls which composed 64 pairs.

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 ROT2015 tabular method helped to identify threats and controls of a slightly better quality than the graphical one. In UTM2016 both methods helped to produce same quality results. For both experiments the TOST results confirmed *the equivalence of two methods 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 ROT2015 the participants reported better perception of the tabular model over the graphical one for PEOU and PU. Such difference in mean was lower than our TOST practical significance threshold of



The figures report experts overall quality assessment of the threats and controls identified for four security tasks in ROT2015 (left) and UTM2016 (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 in SG2014 study and we did not split groups into "good" and "bad".

Fig. 1: Experts assessment by methods and experiments

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
ROT2015	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
UTM2016	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	$2.4 \cdot 10^{-7}$

$\delta = \pm 0.6$. TOST failed to reject the hypotheses about non-equivalence between two methods for PEOU and PU. In UTM2016 the perception of the graphical method significantly increased comparing to ROT2015. So, the two methods have equivalent PEOU and PU which confirmed by TOST results.

6 Retrospective Analysis

In the previous studies (SG2013 and SG2014) we compared graphical method CORAS with different tabular methods. In SG2013 as a tabular method we chose SREP [22] 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

Table 4: Average perception of tabular and graphical SRA methods

ROT2015 results showed that the participants reported higher PEOU and PU for the tabular method than for the graphical one. However, TOST results did not reveal any equivalence of two methods and Wilcoxon results did not confirm stat. sig. of the difference. UTM2016 results revealed that two methods are equivalent with respect to PEOU (stat. sig. 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
ROT2015	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
UTM2016	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

on security experts assessment of the final reports quality. In SG2014 we used tabular method from industry proposed by EUROCONTROL, SecRAM. The participants individually conducted SRA of two tasks from SmartGrid scenario using both methods.

In these experiments we followed the approach by Opdahl and Karpati [25] and used the number of threats and controls identified using a method as a measure of the actual efficacy. Thus, we cannot compare current results with the results from [14,16], 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 supports findings of [14]. For good groups TOST failed to reject the hypothesis about non-equivalence in mean PEOU (p= 0.25) and PU (p= 0.27). For all groups TOST results confirmed the equivalence of two methods w.r.t. mean PEOU ($p = 0.051$) and PU ($p = 0.003$).

The retrospective analysis of SG2014 for all participants revealed: *a)* 10% significantly better mean PEOU in favor of graphical method (MW $p = 0.06$) and *b)* 10% significant equivalence of two methods in mean PU (TOST $p = 0.08$). For good participants TOST failed to reject hypothesis about non-equivalence of two methods in mean PEOU ($p = 0.85$) and PU ($p = 0.43$).

For SG2014 the difference between the results for the perception reported in [16] and the results of the retrospective analysis can be due to the different data collection approach which is discussed in Section 3.

The differences between presented experiments can be due to the changes in treatments. In SG2013 textual and graphical methods have quite clear processed and textual method is good in security controls identification, while graphical one is better for threats identification (see [16, Table III]). In SG2014 the textual method has less clear process which led to better PEOU in favor of graphical method, while both methods have similar PU. In the current experiments (ROT2015 and UTM2016) both textual and graphical method have similar PEOU as they provide clear process. The difference in PU between ROT2015 and UTM2016 can be explained by the change of application scenario. In ROT2015

we used the ROT scenario that was designed by the same organization which designed tabular method and security catalogues. Possibly this combination is a "good fit" which led to better perception of the tabular method. In UTM2016 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.

7 Threats to Validity

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.

Another possible factor that could make the difference between two experiment is changes in the feedback process between experiments. In UTM2016 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 ROT2015 such feedback on the warm-up exercise was not provided. Also in ROT2015 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 UTM2016 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. To mitigate this threat we report and analyze the two experiments separately.

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 [3]. The questionnaire includes 8 questions about PEOU and 9 questions about PU, which were adapted from [14,16].

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 threat to 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 [2]. 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 [33,10]. 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.

8 Discussion

If we consider the threats to validity sufficiently mitigated we obtained the following results:

RQ1 Tabular and graphical methods are equally good w.r.t actual efficacy (i.e. quality of identified threats and security controls).

RQ2 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.

Implication for research. The research community can benefit from the following results of our work:

Equivalence test. Many works in Empirical Software Engineering to compare different treatments look for the difference between them and use standard statistical tests (e.g. t-test, Mann-Whitney, Wilcoxon, and etc.) However, they do not define the range that is sufficient to proof the difference between treatments. In our study for values of 5-item scale we used $\delta = \pm 0.6$ meaning that the difference between treatments A and B is statistically significant if, for example, mean quality delivered by A is 2.8 and 3.5 for treatment B. To test for the equivalence between our treatments we used TOST approach.

Actual efficacy: Quantity vs. Quality. The investigation of the full application of security risk assessment requires more thorough tool to measure the actual efficacy of a method. In the first experiments on the security methods comparison we measured actual efficacy of methods in terms of number of identified threats and controls. However, this approach is not precise as, first, it is important to identify the most critical threats and provide effective mitigations to them rather to identify any possible threat and control and, second, the quantitative measure can be biased by use of security catalogues.

Ease of use. Both tabular and graphical methods can help analysts to produce SRA of a similar quality, but if a method does not have a clear process it may affect people's perception how ease to use is the method. This can be tested by another controlled experiment where participants apply same process (e.g. from CORAS method) with tabular and graphical notation, i.e. classical CORAS as an instance of graphical method and tabular CORAS where tables substitute the diagrams.

Learning curve of graphical method is much steep comparing to the tabular one. The participants had more questions during the warm-up illustrative exercise for graphical method then for the tabular one. Our observation of groups' intermediate results showed that even after illustrative exercise many groups had

difficulties to produce correct diagrams. Tabular notation had a few challenges related to understanding the concepts of primary and supporting assets.

Implications for practice. The main implication of our study for practitioners is that both tabular and graphical- based methods can provide *similar support for SRA*. The most important is that method should provide a *clear process* supporting analyst in identification of *a)* major threats specific to the scenario and *b)* effective security controls to mitigate them.

The results of retrospective analysis of the previous experiments supports these findings. In SG2013 study graphical and tabular methods have similar PEOU and PU as both methods have clear process. In contrast, in SG2014 study the graphical method has higher PEOU than the tabular one because graphical method has significantly clearer process comparing to the tabular method.

Also an important role plays the *fit between SRA components*, i.e. that method and security catalogues are appropriate to the domain of the scenario. In ROT2015 experiment we observed slightly better participants' PEOU and PU, but the results failed to reveal any statistically significant equivalence nor difference between two methods in these variables. At the same time, in UTM2016 tabular and graphical methods were found to be statistically equivalent in terms of participants' PEOU and PU. The possible explanation is that tabular method, scenario, and catalogues for ROT2015 were designed by the same organization and became a "good fit", while in UTM2016 application scenario was changed to UATM scenario by NASA that might be "not a good fit" to same method and catalogues.

9 Conclusion

This paper reported the results of two controlled experiments on comparison of graphical and tabular methods for security risk assessment. The experiments involved 35 and 48 MSc students enrolled to Security Engineering course at Fall 2015 and 2016 at the University of Trento.

In this paper we studied how similar are security methods w.r.t. actual and perceived efficacy. For quality/perception value on 5-item Likert scale we defined the equivalence range $\delta = \pm 0.6$. 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 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.

Acknowledgment

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

References

1. Caralli, R., Stevens, J., Young, L., Wilson, W.: Introducing OCTAVE allegro: Improving the information security risk assessment process. Tech. rep., Software Engineering Institute, Carnegie Mellon University (2007)
2. Carver, J.C., Jaccheri, L., Morasca, S., Shull, F.: A checklist for integrating student empirical studies with research and teaching goals. Empirical Software Engineering 15(1), 35–59 (2010)
3. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quart. pp. 319–340 (1989)
4. Deng, M., Wuyts, K., Scandariato, R., Preneel, B., Joosen, W.: A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. Req. Eng. 16(1), 3–32 (2011)
5. Food and Drug Administration: Guidance for industry: Statistical approaches to establishing bioequivalence (2001)
6. Giorgini, P., Massacci, F., Mylopoulos, J., Zannone, N.: Modeling security requirements through ownership, permission and delegation. In: Proc. of RE 2005. pp. 167–176. IEEE (2005)
7. de Gramatica, M., Labunets, K., Massacci, F., Paci, F., Tedeschi, A.: The Role of Catalogues of Threats and Security Controls in Security Risk Assessment: An Empirical Study with ATM Professionals. In: Proc. of REFSQ 2015. Lecture Notes in Computer Science, vol. 9013, pp. 98–114. Springer (2015)
8. Haley, C., Laney, R., Moffett, J., Nuseibeh, B.: Security requirements engineering: A framework for representation and analysis. IEEE T. Software Eng. 34(1), 133–153 (2008)
9. Hernan, S., Lambert, S., Ostwald, T., Shostack, A.: Threat modeling-uncover security design flaws using the stride approach. MSDN Magazine-Louisville pp. 68–75 (2006)
10. Höst, M., Regnell, B., Wohlin, C.: Using students as subjects: A comparative study of students and professionals in lead-time impact assessment. Empirical Softw. Engg. 5(3), 201–214 (Nov 2000)
11. Karpati, P., Redda, Y., Opdahl, A.L., Sindre, G.: Comparing attack trees and misuse cases in an industrial setting. Inform. Soft. Tech. 56(3), 294–308 (2014)
12. Kopardekar, P.H.: Unmanned aerial system (UAS) traffic management (UTM): Enabling low-altitude airspace and UAS operations. Tech. rep. (2014)
13. Kopardekar, P.H.: Revising the airspace model for the safe integration of small unmanned aircraft systems. Tech. rep. (2015)
14. Labunets, K., Massacci, F., Paci, F., Tran, L.M.S.: An Experimental Comparison of Two Risk-Based Security Methods. In: Proc. of ESEM 2013. pp. 163–172. IEEE (2013)

15. Labunets, K., Paci, F., Massacci, F., Ragosta, M., Solhaug, B.: A First Empirical Evaluation Framework for Security Risk Assessment Methods in the ATM Domain. In: Proc. of SIDs 2014. SESAR (2014)
16. Labunets, K., Paci, F., Massacci, F., Ruprai, R.: An experiment on comparing textual vs. visual industrial methods for security risk assessment. In: Proc. of EmpiRE Workshop at RE 2014. pp. 28–35. IEEE (2014)
17. Landoll, D.J., Landoll, D.: The security risk assessment handbook: A complete guide for performing security risk assessments. CRC Press (2005)
18. Li, T., Horkoff, J.: Dealing with security requirements for socio-technical systems: A holistic approach. In: Proc. of CAiSE 2014. pp. 285–300. Springer (2014)
19. Lund, M.S., Solhaug, B., Stølen, K.: A guided tour of the CORAS method. In: Model-Driven Risk Analysis, pp. 23–43. Springer (2011)
20. Maiden, N., Robertson, S., Ebert, C.: Guest editors' introduction: Shake, rattle, and requirements. IEEE Software 22(1), 13 (2005)
21. Massacci, F., Paci, F.: How to select a security requirements method? a comparative study with students and practitioners. In: Proc. of NordSec 2012. pp. 89–104. Springer (2012)
22. Mellado, D., Fernández-Medina, E., Piattini, M.: Applying a security requirements engineering process. In: Proc. of ESORICS 2006. pp. 192–206. Springer (2006)
23. Meyners, M.: Equivalence tests – a review. Food quality and preference 26(2), 231–245 (2012)
24. Mouratidis, H., Giorgini, P.: Secure tropos: a security-oriented extension of the tropos methodology. Int. J. Inform. Syst. Model. Design 17(02), 285–309 (2007)
25. Opdahl, A.L., Sindre, G.: Experimental comparison of attack trees and misuse cases for security threat identification. Inform. Soft. Tech. 51(5), 916–932 (2009)
26. Scandariato, R., Wuyts, K., Joosen, W.: A descriptive study of Microsoft's threat modeling technique. Req. Eng. pp. 1–18 (2014)
27. Schuirmann, D.: On hypothesis-testing to determine if the mean of a normal-distribution is contained in a known interval. In: Biometrics. vol. 37, pp. 617–617. International Biometric Soc (1981)
28. SESAR: ATM Security Risk Assessment Methodology. SESAR WP16.2 ATM Security (February 2003)
29. Stålhane, T., Sindre, G.: Safety hazard identification by misuse cases: Experimental comparison of text and diagrams. In: Proc. of MODELS 2008. pp. 721–735 (2008)
30. Stålhane, T., Sindre, G.: Identifying safety hazards: An experimental comparison of system diagrams and textual use cases. In: Proc. of BPMDS 2012. vol. 113, pp. 378–392 (2012)
31. Stålhane, T., Sindre, G.: An experimental comparison of system diagrams and textual use cases for the identification of safety hazards. Int. J. Inform. Syst. Model. Design 5(1), 1–24 (2014)
32. Stålhane, T., Sindre, G., Bousquet, L.: Comparing safety analysis based on sequence diagrams and textual use cases. In: Proc. of CAiSE 2010. vol. 6051, pp. 165–179 (2010)
33. Svahnberg, M., Aurum, A., Wohlin, C.: Using students as subjects - an empirical evaluation. In: Proc. of ESEM 2008. pp. 288–290. ACM (2008)
34. Theilmann, C.A.: Integrating Autonomous Drones into the National Aerospace System. Ph.D. thesis, University of Pennsylvania, PA, US (April 2015)
35. Van Lamsweerde, A.: Goal-oriented requirements engineering: A guided tour. In: Proc. of RE 2001. pp. 249–262. IEEE (2001)
36. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering. Springer (2012)