An Experiment on Comparing Textual vs Visual Industrial Methods for Security Risk Assessment

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

Raminder Ruprai
National Grid, UK
Email: raminder.ruprai@nationalgrid.com

***Abstract*—Many security risk assessment methods have been proposed both from academia and industry. However, little empirical evaluation has been done to investigate how these methods are effective in practice. In this paper we report a controlled experiment that we conducted to compare the *effectiveness* and participants' *perception* of visual versus textual methods for security risk assessment used in industry. As instances of the methods we selected CORAS, a method by SINTEF used to provide security risk assessment consulting services, and SecRAM, a method by EUROCONTROL used to conduct security risk assessment within air traffic management. The experiment involved 29 MSc students who applied both methods to an application scenario from Smart Grid domain. The dependent variables were *effectiveness* of the methods measured as number of specific threats and security controls identified, and *perception* of the methods measured through post-task questionnaires based on the Technology Acceptance Model. The experiment shows that while there is no difference in the actual effectiveness of the two methods, the visual method is better perceived by the participants.**

***Index Terms*—controlled experiment, security risk assessment methods, technology acceptance model**

I. Introduction

Many security risk assessment methods, frameworks and standards exist - ISO 27005 [1], NIST 800-30 [2], STRIDE [3], CORAS [4], SREP [5] - but they all face similar problems in practice. The security risk assessment process looks easy on paper - but it can turn into a complex and daunting task.

Despite the crucial role that security risk assessment plays in building secure software systems, only few security engineering papers [6], [7], [8], [9], [10], [11] investigated which methods work better to identify threats and security controls and why. Most of the papers just report proofs of concept discussions based on toy examples. In fact, evaluation of security risk assessment method is challenging because it includes a number of confounding variables: the type of training received (e.g. all papers on the ISACA journal report methods applications by the method's expert), the previous expertise (students vs. practitioners is a key distinction here), the time allocated to the task, and the presence of three essential steps of the analysis (assets, threats and security measures identification depends on each other) so if one is badly performed the others may be poor as well.

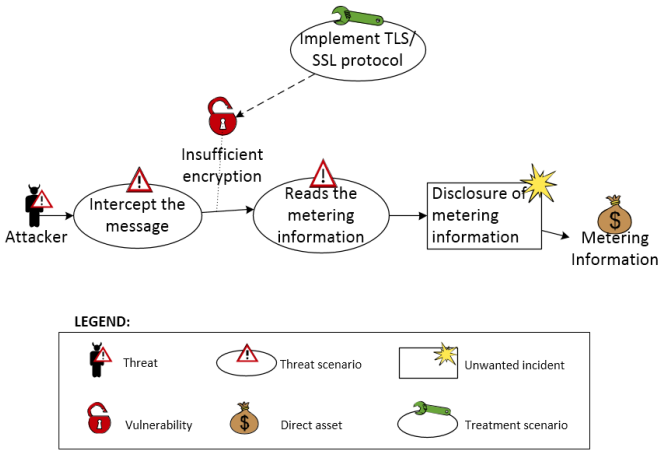
In this paper we report an experiment we conducted to compare *actual effectiveness*, and *perception* of visual versus textual methods for security risk assessment used in industry. We selected CORAS [4] and EUROCONTROL SecRAM [12] as in instances of visual and textual methods respectively. CORAS is a *visual method* whose analysis is supported by a set of diagrams that represent assets, threats, risks and treatments. In contrast, SecRAM is a method used by EUROCONTROL to conduct security risk assessment in the air traffic management domain which mainly uses tables to document the assessment results. We involved 29 participants: 15 students of the MSc in Computer Science and 14 students of the EIT ICT LAB MSc in Security and Privacy of the University of Trento. Each participant applied both methods to identify threats and security controls for different security facets (Network security and Database/Web application security) of a Smart Grid application scenario. The experiment was complemented with participants' interviews to explain possible differences between the two methods.

The main findings on effectiveness are that there is no difference in the number of threats and security controls identified with each method. With respect to participants' perception, we found that the visual method is preferred over the textual one with statistical significance.

We also compared the results of this experiment with the previous [11] where we investigated the differences in actual effectiveness and perception of visual and textual methods from academia. The main difference in the two experiments is that participants had to work in group rather than individually while the application scenario was exactly the same in both experiments. This experiment only confirms the results from the first experiment on perception: the visual method has higher participants' perception than the textual one.

II. Related Work

The few papers [6], [7], [8], [9], [10], [13], [11] that attempted to evaluate if security risk assessment methods work in practice adopted the Method Evaluation Model (MEM) [14] which provides constructs to measure methods success. For example, Opdahl and Sindre [6] carried out two controlled experiments (28 and 35 students) to compare two methods for threats identification, namely attack trees and misuse cases. In [10] Opdahl and colleagues repeated the experiment with industrial practitioners. Both experiments showed that attack trees help to identify more threats than misuse cases. Similar controlled experiments with students were reported by Stålhane et al. in [15], [8], [9], [7] where misuse cases are



(a) CORAS - Threat Diagram

Threat Agent	Asset Attacked	Attack Likelihood	Justification
Compromised MPO	SM, EMS	Probable	Can use message replay attack and access customer data
Malicious attacker	EMS, HAN, SA, S&C	Probable	By Eavesdropping and Sniffing on the HAN, can use DOS attack to deny availability of HAN, Hacking the EMS and tampering the S&C and accessing the SA

(b) SECRAM - Threat Agent Table

Fig. 1: Examples of Visual (CORAS) and Textual (SecRAM) Methods' Artifacts Generated by Participants.

compared with other approaches for safety and security. In [15] Stålhane et al. reported an experiment with 42 students where they compared misuse cases to Failure Mode and Effects Analysis (FMEA) to analyze use cases. They found that misuse cases are better than FMEA for analyzing failure modes related to user interactions. In a similar setting [8] the authors compared misuse cases based on use-case diagrams to those based on textual use cases. The results of the experiment with 52 students showed the that textual use cases produces better results due to more detailed information. Massacci and Paci [13] reported the results of the eRISE challenge where methods from academia for security analysis were applied by both practitioners and students. The challenge revealed that threat-based methods performs better for security analysis. More recently, Labunets et al. [11] conducted a controlled experiment with 28 MSc students to compare two types of security risk assessment methods, visual (CORAS) and textual (SREP) methods. The participants worked in groups and applied methods to four security facets from Smart Grid application scenario. The results showed that visual methods are more effective in identifying threats and better perceived than the textual ones. We also conducted a controlled experiment with MSc students to compare visual (CORAS) and textual (SecRAM) methods. However, in our experiments participants worked individually to apply both methods to two security facets from Smart Grid scenario.

As in [6], [7], [8], [9], [10], [11], in our experiment we also used MEM as basis to compare textual and visual methods for security risk assessment. However, in [6], [7], [8], [9], [10] actual effectiveness of the methods was evaluated based on the number of artifacts identified by the participants. In contrast, as in [11], in our experiment we determine effectiveness based on the quality of threats and controls identified by the participants because a method is effective if it produces good results. To avoid bias in evaluation, we asked two external experts to assess the quality of the threats and controls produced by the participants. In addition, the experiments reported in [6], [7], [8], [9] had a short duration (less than two hours) and this may have introduced bias in the evaluation of methods because subjects did not have enough time to understand the application scenario and to fully apply the methods under evaluation. Further, since the time for the execution of these experiments was short, the methods have been applied to toy scenarios and the results might not generalize to real-world cases. In our experiment, the participants received training on the application scenario and the methods of the duration of two hours each. They also had more than two weeks to apply the methods to the application scenario rather than just two hours. In addition, the participants applied the methods to a real industrial application scenario.

III. Research Method

This section describes the design of the performed experiment, following the guidelines by Wohlin et al. [16].

A. Research Questions

The *goal* of the experiment was to compare visual and textual methods for security risk assessment with respect to how successful they are in identifying threats and security controls. For this purpose we adopted as dependent variables the success constructs defined in the Method Evaluation Model (MEM) proposed by Moody [14]: *effectiveness*, *perceived easy of use* (PEOU), *perceived usefulness* (PU), and *intention to use* (ITU). Therefore, we specified the following research questions:

RQ1 *Is the effectiveness of the methods significantly different between the two types of methods?*

RQ2 *Is the participants' overall perception of the method significantly different between the two type of methods?*

RQ3 *Is the participants' perceived usefulness of the method significantly different between the two type of methods?*

Variable	Scale	Means	Distribution
Gender	Sex		79% were male; 21%were female
Age	Years	25.72	48% were 21-24 years; 41% were 25-29; 10% were 30-40
Education Length	—"—	4.28	66% had <5 years; 17% had 5 years; 17% had >5 years
Work Experience	—"—	2.46	31% had no experience; 31% had < 2 years; 28% had 3-5 years; 10% had >6 years
Level of Expertise in Security Technology	1(Novice)-5(Expert)	2.31	28% novices; 28% beginners; 10% competent users; 31% proficient users; 3% experts
Level of Expertise in Security Regulation and Standards	—"—	1.86	45% novices; 17% beginners; 7% competent users; 31% proficient users
Level of Expertise in Privacy Technology	—"—	2.10	31% novices; 34% beginners; 28% competent users; 7% proficient users
Level of Expertise in Privacy Regulation	—"—	1.90	48% novices; 24% beginners; 7% competent users; 21% proficient users
Level of Expertise in RE	—"—	2.31	24% novices; 34% beginners; 14% competent users; 28% proficient users

TABLE I: Demographic Statistics

RQ4 *Is the participants' perceived ease of use of the method significantly different between the two type of methods?*

RQ5 *Is the participants' intention to use the method significantly different between the two type of methods?*

We translated research questions $RQ1 - RQ5$ into a list of null hypotheses to be statistically tested. We do not list them here due to the lack of space. To answer $RQ1$ we measured methods' *actual effectiveness* by counting the number of threats and security controls identified with each method application and we asked two external security expert to assess their quality. $RQ2$-$RQ5$ was answered by administering a post-task questionnaire inspired to the Method Evaluation Model (MEM) [14]. To gain better understanding *why there is a difference in methods effectiveness and perception* we conducted individual interviews with the participants.

B. Methods Selection

As instance of the visual method we chose CORAS [4], a model-driven method designed by SINTEF, a research institution in Norway, which uses it to provide consulting services. It consists of three tightly integrated parts: a method for risk analysis, a language for risk modeling, and a tool to support the risk analysis process. The risk analysis in CORAS is a structured and systematic process which uses diagrams (see Figure 1a) to document the results. The steps are based on ISO 31000 for risk management [17]: context establishment, risk analysis (that identifies assets, unwanted incidents, threats and vulnerabilities), and risk treatments. As instance of textual method we selected SecRAM [12], an industrial method used by EUROCONTROL to conduct security risk assessment in the air traffic management domain (ATM). SecRAM supports the security risk assessment process for a project initiated by an air navigation service provider, or ATM project, system or facility. It provides a systematic approach to conduct security risk assessment which consists of five main steps: defining the scope of the system, assessing the impact of a successful attack, estimating the likelihood of a successful attack, assessing the security risk to the organization or project, and defining and agreeing a set of management options. As shown in Figure 1b) tables are used to represent the results of each step's execution.

C. Domain Selection

We selected an application scenario from Smart Grid domain. The Smart Grid is an electricity network that uses information and communication technologies to optimize the distribution and transmission of electricity from supply points to consumers. The application scenario was focused on gathering of metering information from smart meters located in private households and its communication to electricity suppliers for billing purposes.

D. Demographics

The participants were recruited among MSc students enrolled in the Security Engineering course at the University of Trento. Table I presents descriptive statistics about the participants. Most of participants (69%) reported that they had at least 2 years of working experience while the remaining said they had no working experience. With respect to knowledge in privacy technologies and regulations, most of the participants had limited expertise. In contrast, they reported an extensive general knowledge of both security technologies and regulations and standards. Participants also reported good general knowledge in requirements engineering.

E. Experimental Design

We chose a within-subject design where all participants apply both methods to ensure a sufficient number of observations to produce significant conclusions. In order to avoid learning effects, the participants had to identify threats and security controls for different security facets of a Smart Grid application scenario. The security facets included Network Security (Network) and Database/Web Application Security (DB/WebApp). For example, for Network Security facet, participants had to identify threats like man-in-the-middle attack or DoS attack and to propose security controls to mitigate them.

Participants were randomly assigned to treatments: one half of participants applied first the visual method to Network Security facet and then the textual method for the Database/Web Application Security facet, while the other half applied the methods in the opposite order.

F. Experimental Procedure

The experiment was performed during the Security Engineering course held at University of Trento from September 2013 to January 2014. The experiment was organized in three main phases:

Training. Participants were given a 2 hours tutorial on the Smart Grid application scenario and a 2 hours tutorial on visual

and textual methods. Then, participants were administered a questionnaire to collect information about their background and their previous knowledge of other methods, and they were assigned to facets based on the experimental design.

Application. Once trained on the Smart Grid scenario and the methods, the participants had to repeat the application of the methods on two different facets: Network and DB/WebApp. For each facet participants:

- Attended a two hours lecture on the threats and possible security controls specific to the facet but not concretely applied to the scenario.
- Had 2,5 weeks to apply the assigned methods to identify threats and security controls specific for the facet.
- Gave a short presentation about the preliminary results of the method application and received feedback.
- Had one week to deliver an intermediate report to get feedback.

At the end of the course in mid January 2014 each participants submitted a final report documenting the application of the methods on the two facets.

Evaluation. In this phase the participants provide feedback on the methods through questionnaires and interviews. After each application phase participants answered an on-line post-task questionnaire to provide their feedback about method. The post-task questionnaires were inspired by the Technology Acceptance Model (TAM) [18]. To prevent participants from "auto-pilot" answering, 15 out of 31 questions were given with the most positive response on the left and the most negative on the right. In addition, after final report submission each participant was interviewed for half an hour by one of the experimenters to investigate which are the advantages and disadvantages of the methods. The interview guide contained open questions about the overall opinion of the methods, whether the methods help in identification of threats and security controls and about methods' possible advantages and disadvantages. The interview questions were the same for all the interviewees. The interview guide and the post-task questionnaire are reported in [19].

IV. Quantitative Analysis

In this section we report the results from the analysis of the final reports delivered by the participants and of the participants' answers to the post-task questionnaires.

A. Quality of Results

Since a method is effective based not only on the quantity of results, but also on the quality of the results that it produces, we asked two domain experts to independently evaluate each individual report. To evaluate the quality of threats and security controls experts used a four item scale: *Unclear* (1), *Generic* (2), *Specific* (3) and *Valuable* (4). In terms of the final assessment we observed that:

1) the experts marked bad participants the same way,
2) they consistently marked moderately good students,
3) a couple of students were border line. In other words their threats and controls were neither definitely good nor bad.
4) they had a different evaluation only for 3 out of 29 students. This may be explained by the different expertise of the domain experts: more management and seniority of one expert, more operational and junior other expert.

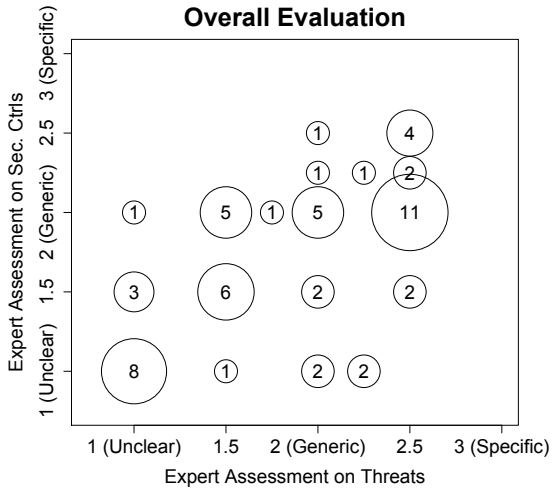


Fig. 2: Overall experts assessment of threats and security controls for the two facets.

In order to validate whether the difference in experts' evaluation is statistically significant we run Wilcoxn non-parametric paired test. The results show that there is no statistically significant differences in the evaluations of two experts ($p = 0.58$).

Figure 2 illustrates the average of the evaluation of the two experts for all participants. As each participant applied one of the methods on both facets, there are 58 method applications in total. The number inside each bubble denotes the number of method applications which achieved a given assessment for threats (reported on x-axis) and security controls (reported on y-axis). There were 24 method applications that generated some specific threats and/or security controls. The remaining method applications delivered unclear and/or generic threats and security controls.

We evaluated the actual effectiveness of methods based of the number of *specific* threats and security controls. In what follows, we will compare the results of all methods' applications with the results of those applications that produce specific threats and security controls.

B. Reports Analysis

To assess the effectiveness of visual and textual methods, we reviewed the final reports delivered by the participants to count the number of identified threats and security controls.

As the design of our experiment is two factor block design (the method and the facet), we could use the two-way ANOVA test or Friedman test (non-parametric analog of ANOVA) to analyze the number of threats and security controls identified with each method and within each facet. To select a right test we checked whether our samples satisfy ANOVA's assumptions: *a)* observations independence, *b)* homogeneity of

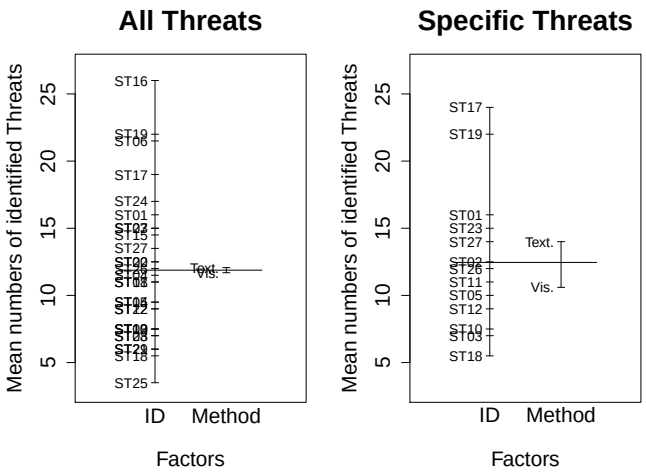


Fig. 3: Means of all identified threats (left) and specific ones (right).

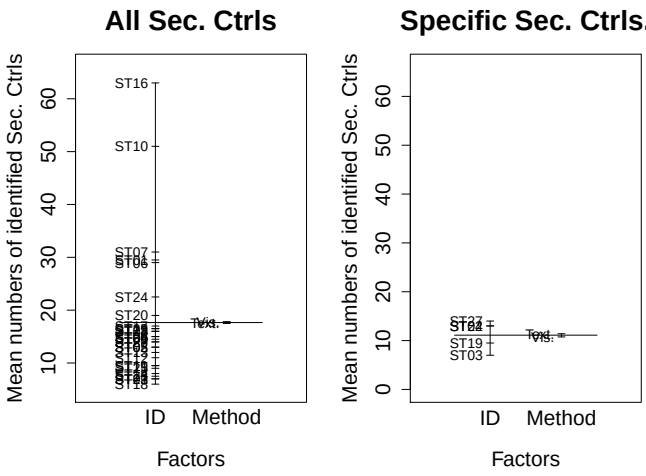


Fig. 4: Means of identified alls security controls (left) and specific ones (right).

variance *c)* normality of distribution of samples. We set the significance level $\alpha = 0.05$.

Observation Independence. We have *observation independence* by design because participants' worked individually. This gave us independence within sample and mutual independence within sample as the facets were different.

Homogeneity of Variance. We checked the homogeneity of variance with Levene's test. This test returned p equal to 0.27 for threats and 0.52 for security controls. Therefore, we can assume homogeneity of variance for our samples.

Distribution Normality. To check this assumption we used Shapiro-Wilk normality test. This test returned $p = 0.01$ for threats and 0.93 for security controls. So, normality assumption holds only for the security controls.

Therefore, we could use Friedman test to analyze the difference in the number of threats and ANOVA test for security controls. However, since we also considered *specific* results, we had unbalanced samples because some participants produced specific threats and security controls for the application of one method while for the other method they did not. Therefore, we used the analog of Friedman test, Skilling-Mack test [20], that can work with unbalanced samples for the analysis of the difference in the *number of threats*, and ANOVA test with Type II of Sum of Squares [21] for the analysis of the difference in the number of *security controls*.

Figure 3 shows that the textual method is better than the visual one in identifying threats. But the results of the statistical tests did not show any significant differences in the number of threats among both all (Friedman test returned *p-value* = 0.57) and specific threats (Skillings–Mack test returned *p-value* = 0.17).

In contrast, Figure 4 shows that the visual and textual method produce the same number of security controls. This is attested also by the results of statistical tests which showed there was no statistically significant difference in the number of security controls of any quality (Friedman test returned *p-value* = 0.57) and specific security controls (ANOVA test returned *p-value* = 0.72).

We also found that there is no statistically significance difference in the number of threats and controls identified by the participants within each security facet.

C. Questionnaire Analysis

The post-task questionnaires was analyzed to identify the difference in participants perception of two methods. Before conducting analysis all responses were reverted to 5 being the best. The questions were formulated in opposite statements format with answers on a 5-point Likert scale. We compare the answers of all participants with the answers of those participants whose methods applications produced specific threats and/or security controls (denoted as *good subjects* in what follows). We analyzed the answers of all participants with Wilcoxon test since the data are ordinal and the responses of participants are paired. Instead, we used Mann-Whitney (MW) test to analyze the answers of participants who produced specific results because some observations were unpaired. Since MW test requires homogeneity of variance of samples, we checked this assumption.

Homogeneity of Variance. The Levene's test revealed that in general the variances of our samples are equal ($p = 0.95$). However, there is no equal variance for responses on overall PEOU of method ($p = 0.036$). Thus, we could not consider the results of MW test of this category as valid.

Table II presents the results of questionnaires' analysis. For each question, the table reports to which perception variable the question refers to (PEOU, PU, ITU), the mean of the answers, and Z statistics returned by the test and the level of statistical significance based on the p-value returned by the test. The level of statistical significance is specified by • (p<0.1), or * (* p<0.05, ** p<0.01, *** p<0.001).

Perceived Ease of Use. The visual method is better than the textual with respect to overall PEOU and the difference

		All subjects				Good subjects		
Q	Type	Mean		Z_W	Z_{MW}	Mean		Z_{MW}
		Tex	Vis			Tex	Vis	
1	PU	3.72	4.14	-2.11 *	-1.63	3.5	4.2	-1.27
2	Control	3.72	3.9	-0.36	-0.1	3.8	4.1	-0.36
3	Control	3.79	4.07	-0.85	-0.75	3.6	4.1	-0.75
4	PU	3.14	3.83	-2.41 *	-2.15 *	3.3	3.8	-0.83
5	PU	3.17	3.59	-1.58	-1.53	3.3	3.4	-0.08
6	PEOU	2.93	3.9	-2.9 **	-2.84 **	3.1	3.9	-1.42
7	PEOU	2.93	3.69	-2.58 **	-2.5 *	2.9	3.6	-1.34
8	PU	3.55	4	-1.61	1.69 •	3.4	3.9	-0.9
9	PEOU	2.79	3.79	-3.33 ***	-2.98 **	2.6	3.7	-1.84 •
10	PU	3	3.83	-2.5 *	-2.63 **	3.1	3.9	-1.59
11	PU	2.9	3.48	-2.22 *	-1.89 •	3.2	3.5	-0.39
12	PU	3.17	3.1	0.4	0.09	3	3.1	-0.12
13	PU	3.03	2.97	0.23	0.06	2.7	2.6	0.27
14	PU	3.17	3.45	-1.35	-0.73	3.4	3.6	-0.27
15	ITU	3.07	3.69	-1.78 •	-1.97 •	3.1	3.4	-0.51
16	ITU	3.28	3.45	-0.64	-0.66	3.3	3.7	-0.8
17	Control	2.86	3.69	-3.16 **	-3.42 ***	2.6	3.6	-1.96 •
18	Control	3	3.62	-2.67 **	-2.66 **	3	3.5	-1.02
19	ITU	3.1	3.76	-2.17 *	-2.14 *	3	3.8	-1.44
20	ITU	3.17	3.59	-1.3	-1.39	3.2	3.6	-0.64
21	Control	3.34	2.69	2.01 *	1.89 •	2.9	2.8	0.15
22	PU	3.1	3.38	-1.03	-0.89	2.7	3.5	-1.44
23	ITU	3.1	3.55	-1.68 •	-1.53	3.2	3.6	-0.52
24	ITU	3.07	3.38	-1.08	-1.05	3.1	3.5	-0.55
25	PU	3.28	3.69	-2.15 *	-1.63	3	3.7	-1.21
26	PU	3.07	3.52	-1.53	-1.47	3	3.2	-0.39
27	PEOU	3.03	3.9	-2.78 **	-2.78 **	2.9	3.9	-1.92 •
28	ITU	3.14	3.48	-1.33	-1.22	3.1	3.6	-0.92
29	ITU	3.21	3.28	-0.25	-0.39	3.2	3.3	-0.32
30	PEOU	2.93	3.55	-2.8 **	-1.91 •	2.5	3.4	-1.54
31	PEOU	3.14	3.86	-2.15 *	-2.07 *	2.7	3.8	-2.14 *
PEOU		2.96	3.78	-6.61 ***	-6.16 ***	2.78	3.72	-4.21 ***
PU		3.19	3.58	-5.18 ***	-4.56 ***	3.13	3.53	-2.39 *
ITU		3.14	3.52	-3.67 ***	-3.67 ***	3.15	3.56	-2.05 *
Total		3.16	3.61	-8.93 ***	-3.67 ***	3.08	3.59	-5.24 ***

• - p-value <0.1, * - p <0.05, ** - p <0.01, *** - p <0.001

TABLE II: Mann-Whitney Test of Responses of All and Good Subjects

PEOU Category	Vis.	Text.	Total
Positive Aspects			
Clear Process	28	18	46
Visual summary	43		43
Time effective	7	16	23
Easy to Understand	18		18
Worked examples	12	4	16
Easy for Customer	13	2	15
Total Pos PEOU	121	40	161
Negative Aspects			
Time consuming	36	7	43
Unclear Process	4	28	32
Primitive Tool	30		30
Poor worked examples	2	27	29
Not easy to Use	6	18	24
Redundant Steps	19	4	23
No Evolution Support	15	2	17
Not easy to Understand	3	11	14
Total Neg PEOU	115	97	212
Total PEOU	236	137	373

PU Category	Vis.	Text.	Total
Positive Aspects			
Help in Identifying Threats	39	18	57
Help in Identifying Security Controls	22	16	38
Help to Model	10	2	12
Total Pos PU	71	36	107
Negative Aspects			
No Help in Identifying Security Controls	9	16	25
No Tool Support		21	21
Visual Complexity	17		17
Total Neg PU	26	37	63
Total PU	97	73	170

Other Category	Vis.	Text.	Total
Positive Aspects			
Catalog of Sec. Controls	23	31	54
Catalog of Threats	30	29	59
Total Pos Other	53	60	113

TABLE III: Positive and Negative Aspects Influencing Method Perception

is statistically significant (for good subjects MW returned: $Z(good)_{MW}$ = -4.21, p = $2 * 10^{-5}$, es = 0.38). But we cannot rely on this result because homogeneity of variance assumption is not met.

Perceived Usefulness. The visual method is better than the textual with respect to overall PU with statistical significance ($Z(good)_{MW}$ = -2.39, p = $1.7 * 10^{-2}$, es = 0.15).

Intention to Use. The visual method is better than the textual with respect to overall ITU with statistical significance ($Z(good)_{MW}$ = -2.05, p = $3.9 * 10^{-2}$, es = 0.16).

Overall Perception. The average of responses shows that participants preferred the visual method over the textual method with statistical significance ($Z(good)_{MW}$ = -5.24, p = $1.4 * 10^{-7}$, es = 0.21).

V. Qualitative Analysis

In this section we report the results of the analysis of individual interviews with participants. The interviews were transcribed and analyzed by two researchers independently using *coding* [22], a qualitative analysis method from grounded theory. The list of core codes was taken from analysis of previous experiments [11], [13].

Table III reports the positive and negative aspects of visual and textual methods that may affect PEOU and PU and *other* aspects that may influence methods' success. For each aspect we report the total number of statements made by participants as relative indicator of its importance. Here we report only the aspects for which at least 10 statements were made by participants.

Perceived Ease of Use. The main aspect influencing PEOU of visual method is that it provides a *visual summary* of the results of the security analysis (29% of the positive statements made by the participants on visual method's PEOU). Examples of these statements are: "*there are many summary diagrams which are useful to summarize what has been done*" and "*the advantage is the visualization*". Another noteworthy positive aspect for visual method's PEOU is that the visual method has *clear process* (19% of positive statements):"*The advantages of CORAS is very clear structure*". Instead, the main aspects that can affect negatively the visual method's PEOU are that it is a *time consuming* method and it has a *primitive tool* (26% of negative statements). As participants indicated "*the diagrams are really time consuming*" and "*first I tried the CORAS tool. And somehow, it was confusing. So, I switched to the Visio*". Another negative aspect for visual method's PEOU is that the process has *redundant steps* (17% of negative statements): "*I think CORAS has some duplications.*".

The main positive aspect for the textual method's PEOU is *time effectiveness* (26% of positive statements): "*I used very little time to do my work*". Instead, there is no consensus among participants about other two aspects: *clear process* and *ease of use*. In fact, participants made a similar number of statements that indicate these aspects as both positive and negative: "*it's quite easy*" (positive statement) and "*it was sometimes a bit confusing how to apply the methodology*" (negative statement).

The main negative aspect (28% of negative statements) impacting textual method's PEOU is related to *poor worked examples* illustrating method application. As participants reported "*the main problem was about the example that it uses - instead of defining in more general way, and you are misguided by this example*".

Perceived Usefulness. There are two main aspects that could positively affect PU of visual method: *help in identifying threats* (55% of positive statements) and *security controls* (31% of positive statements): "*when you're doing a diagram you can actually see the flaw of the actions and it is easy to identify the threats, the attacks*" and "*I find it good for finding some security requirements and risk*". The negative aspect for visual method PU is that visual notation does not scale well for complex scenarios (65% of negative statements): "*these diagrams are getting soon very huge and very complex*".

Similarly, the main positive aspect for textual method PU is that "*it has detailed steps and helps to identify assets, threat agents and management options*" (50% of positive statements). Instead, there is no consensus among participants about the textual method helping in the *identification of security controls*. In fact, they made equal number of positive and negative statements about this aspect. Here are examples of typical statements made by participants about it: "*After we already known that our system description, the vulnerabilities, the threat or agents is easy to identify the control.*" (positive statement) or " *I can't say that they allow you to find the threat, the security control, whatever you want. It's just a framework to help you.*" (negative statement).

The most significant negative aspect mentioned for textual method's PU is the fact there is no software supporting the execution of the steps of the textual method: "*It is needed because it would save half of the time if the table were generated automatically*" (57% positive statements).

Other Relevant Aspects. In participants' interview we also identified other possible aspect influencing methods' success. Participants think that both methods would benefit from availability of catalogs of threats and security controls: "*I think that SecRAM could just employ some catalog by default.*".

VI. Discussion

In this section we present the main findings regarding each of the research questions. Table IV compares them with the findings from the first experiment where we compare visual and textual methods from academia [11].

Methods' effectiveness. As shown in the previous sections, there is no difference in the number of threats and controls identified with each method. Therefore, we can reject the alternative hypotheses $H1.1_A$ and $H1.2_A$. In contrast, in first experiment $H1.1_A$ was accepted: visual method performed better in threats identification. This difference may be due to the change of the textual method: SecRAM could perform better than SREP. Or due to the difference in the experimental design. In the first experiments participants applied each method twice, while in the present experiment there was only one application of the method. The participants of the first experiment might have learnt methods better and produced significant results.

Methods' perception. Participants' *overall perception* is higher for visual than for textual method with statistical significance for all and good participants. Alternative hypothesis $H2_A$ of difference in the overall perception of the two methods is thus upheld. The same result holds for PU and ITU. Thus, the alternative hypotheses $H4_A$ and $H5_A$ can be accepted. However, the hypothesis $H3_A$ remains open because PEOU sample did not meet required test assumptions. Similar results were found in the first experiment. The overall perception and ITU were higher for the visual method, while for PU and PEOU there was no evidence to tell if there was a difference between the two methods.

Qualitative Explanation. The different perception of the method: visual method perceived better than textual one, can be likely explained by the differences between the two methods indicated by the participants during the interviews. Diagrams in visual method help participants to model the system and help in identifying threats and security controls because they give an overview of the possible threats (who initiate the threats), the threat scenarios (possible attacks) and the assets, while the identification of threats in textual method is not facilitated by the use of tables because it is difficult to keep the link between assets and threats and the process is unclear. Also, lower perception of textual method can be explained by a poor worked example illustrating method application, and the unavailability of the software that would help to generate a bulk of tables.

VII. Threats to Validity

In this section we discuss the main types of threats to validity [16]. **Internal validity**. One expected threat to internal validity is related to possible *bias in the tutorials*. Differences in the methods' performance may occur if a method is presented in a better way than the other. In our experiment we limit this threat by giving the same structure and the same duration to the tutorials on textual and visual methods. Finally, *bias in data analysis* was limited by having the participants' reports coded by the authors of the paper independently. In addition, the quality of the threats and security controls identified by each group was assessed by two domain experts external to the experiment.

Construct validity. The main threat to construct validity in our experiment is the design of the research instruments: interviews and questionnaires. The questionnaire was designed following TAM with at least six questions for each of the

TABLE IV: Results of hypothesis testing

Id	Hypotheses	1st experiment	Current experiment
$H1.1_A$	Difference in the number of threats found with visual and with textual method	YES	NO
$H1.2_A$	Difference in the number of security controls found with visual and with textual method	NO	NO
$H2_A$	Difference in the participants preference for visual and textual method	YES	YES
$H3_A$	Difference in the participants perceived ease of use for visual and textual method	MAY BE	MAY BE
$H4_A$	Difference in the participants perceived usefulness for visual and textual method	MAY BE	YES
$H5_A$	Difference in the participants intention to use for visual and textual method	YES	YES

* We re-done statistical analysis on data from the first experiment with Friedman test used in this experiment

independent variables we wanted to measure: *perceived usefulness*, *perceived easy of use*, *intention to use*. Three researchers independently checked the questions included in the interview guide and in the questionnaire: therefore we are reasonably confident that our research instruments measured what we wanted to measure.

Conclusion validity. A main threat to conclusion validity is related to how to evaluate the effectiveness of the methods under evaluation. A method is effective based on the quality of the results that it produces. If we consider just the number of results (e.g., number of threats identified) but not the quality, threats to conclusion validity may arise. To mitigate this threat, we asked two experts in security for Smart Grid to evaluate the results the subjects produced.

External validity. External validity is affected by the objects and the subjects chosen to conduct the experiment. The main threat is related to the *use of students instead of practitioners*. We mitigated this threat by using MSc students enrolled in a course on security engineering. This allowed us to rely on students with the required expertise in security and to ensure that they had the same level of knowledge on the subject. Another threat is the *realism of the experimental environment*. Our experiment had the duration of three months rather than two hours like most of the experiment. This allows us to use a realistically-sized application scenario and thus to generalize our results to real-world cases.

VIII. Conclusions

In this study we compared the effectiveness and the perception of visual versus textual methods for security risk assessment adopted in industry. The main findings on effectiveness are that both methods have similar performance in identification of threats and security controls. With respect to participants' perceived usefulness and intention to use we found that the visual method is preferred over the textual one with statistical significance.

To sum up the intentions for future works, we plan to carry out a replication of this experiment with practitioners in order to generalize our findings. In addition, we will conduct experiments to evaluate the effect that some of the aspects that we identified during interviews have on the effectiveness and perception of the methods.

Acknowledgment

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

References

[1] ISO, "Iso/iec 27005, information technology security techniques - information security risk management," Tech. Rep., 2011.

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

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

[4] M. S. Lund, B. Solhaug, and K. Stolen, "A guided tour of the CORAS method," in *Model-Driven Risk Analysis*. Springer, 2011, pp. 23–43.

[5] D. Mellado, E. Fernández-Medina, and M. Piattini, "Applying a security requirements engineering process," in *Proc. of ESORICS '06*. Springer, 2006, pp. 192–206.

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

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

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

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

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

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

[12] *EATM, ATM Security Risk Assessment Methodology, Edition 1.0*, EUROCONTROL, May 2008.

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

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

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

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

[17] ISO/IEC, *31000:2009 – Risk Management*, 2009.

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

[19] UNITN, "Experiment website," http://securitylab.disi.unitn.it/doku.php?id=seceng-course-exp-2013.

[20] M. Chatfield and A. Mander, "The Skillings–Mack test (Friedman test when there are missing data)," *The Stata Journal*, vol. 9, no. 2, p. 299, 2009.

[21] Ø. Langsrud, "ANOVA for unbalanced data: Use Type II instead of Type III sums of squares," *Statistics and Computing*, vol. 13, no. 2, pp. 163–167, 2003.

[22] A. L. Strauss and J. M. Corbin, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, 1998.