An Experimental Comparison of Two Risk-Based Security Methods

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

***Abstract*—A significant number of methods have been proposed to identify and analyze threats and security requirements, but there are few empirical evaluations that show these methods work in practice. This paper reports a controlled experiment conducted with 28 master students to compare two classes of risk-based methods, visual methods (CORAS) and textual methods (SREP). The aim of the experiment was to compare the *effectiveness* and *perception* of the two methods. The participants divided in groups solved four different tasks by applying the two methods using a randomized block design. The dependent variables were *effectiveness* of the methods measured as number of threats and security requirements identified, and *perception* of the methods measured through a post-task questionnaire based on the Technology Acceptance Model. The experiment was complemented with participants' interviews to determine which features of the methods influence their effectiveness. The main findings were that the visual method is more effective for identifying threats than the textual one, while the textual method is slightly more effective for eliciting security requirements. In addition, visual method overall perception and intention to use were higher than for the textual method.**

Keywords*—*controlled experiment, risk-based methods, technology acceptance model

I. Introduction

Several methods have been proposed to address security concerns during the early phases of the system development life cycle [1]–[6]. However, there has been little empirical evaluation that shows how effective these methods are in practice. With few exceptions [7]–[10], security methods are evaluated by the same researchers who have proposed them. As a consequence, security practitioners are not motivated to adopt new security methods, while researchers do not know how to improve their methods. To address this problem, there is thus the pressing need of conducting empirical evaluations to investigate which methods work better to identify threats and mitigations (i.e., security requirements for later phases) and why.

In this paper, we report a controlled experiment that we conducted to compare two classes of risk-based methods: visual methods and textual methods. As instances of these classes of methods, we have selected CORAS [2] and SREP [1]. CORAS is a *visual method* whose analysis is supported by a set of diagrams that represent assets, threats, risks and treatments. In contrast, SREP is a *textual method* whose artifacts are specified in natural language or in tabular form.

The goal of the experiment was to evaluate the *effectiveness* of visual and textual based methods, and the participants' *perception* of the methods. Hence, the dependent variables were the *effectiveness* of the methods measured as number of threats and security requirements and the participants' *perceived easy of use*, *perceived usefulness* and *intention to use* of the two methods. The independent variable was the *method*. The experiment involved 28 participants: 16 students of the master in Computer Science and 12 students of the EIT ICT LAB master in Security and Privacy. They were divided into 16 groups using a randomized block design. Each group applied the two methods to identify threats and security requirements for different facets of a Smart Grid application scenario (ranging from security management to database security). The experiment was complemented with participants' interviews to gain insights on *why* the methods are effective or they are not.

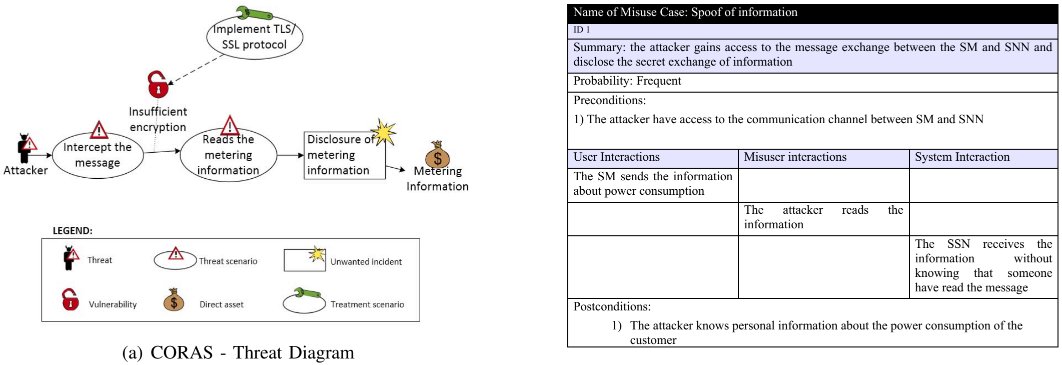
The main findings are that the visual method yields to identify more threats than textual one, while the textual one is slightly better to identify security requirements. The difference in the number of threats identified with the two methods is statistically significant and participants' interviews suggests that this is due to the difference in the artifacts used to model threats. The visual method uses diagrams to represent threats while the textual method uses tables: diagrams help brainstorming on threats and thus yield participants to identify more threats. On the contrary, the difference in the number of security requirements identified with the two methods is not statistically significant. The textual method identified a slightly higher number of security requirements but this is not statistically significant. A possible explanation emerging from the interviews is that process supported by the textual method offers a systematic approach to identify security requirements. In addition, the visual method's overall perception and intention to use are higher than for the textual method.

In the next section we discuss related works (§II) and we present the design and execution of the experiment (§III). The core of the paper reports on the analysis of the participants' reports (§IV), the post-task questionnaire (§V), and the interviews (§VI). Then, we summarize the main findings (§VII) and discuss the threats to validity (§VIII). Finally, we present conclusions and future work (§IX).

II. Related Work

According to a survey conducted in 2009 [11], only 13% of papers reporting research in requirements engineering were based on a case study. In the realm of methods for eliciting threats and security requirements, papers reporting empirical studies are even less frequent. To the best of our knowledge,

978-0-7695-5056-5/13 $26.00 © 2013 IEEE
DOI 10.1109/ESEM.2013.29



(a) CORAS - Threat Diagram

Name of Misuse Case: Spoof of information		
ID 1		
Summary: the attacker gains access to the message exchange between the SM and SNN and disclose the secret exchange of information		
Probability: Frequent		
Preconditions: 1) The attacker have access to the communication channel between SM and SNN		
User Interactions	Misuser interactions	System Interaction
The SM sends the information about power consumption		
	The attacker reads the information	
		The SSN receives the information without knowing that someone have read the message
Postconditions: 1) The attacker knows personal information about the power consumption of the customer		

(b) SREP - Threat Specification using misuse cases

Fig. 1: Examples of Visual (CORAS) and Textual (SREP) Methods' Artefacts.

only two proposals actually compared methods for identifying threats and security requirements [7], [10]. Opdahl et al. [7] have carried out two controlled experiments (with 28 and 35 students, respectively) to compare two methods for threats identification, namely attack trees and misuse cases. They assessed methods' effectiveness, coverage, perceived usefulness, ease of use and intention to use. Similarly, in our experiment, we assessed methods' effectiveness and participants' perception. We also investigated through interviews which are the aspects that influence methods' effectiveness.

Massacci et al. [10] reported the results of eRISE (engineering RIsk and SEcurity Requirements) challenge, a qualitative study conducted in 2011 with 36 practitioners and 13 master students. The aim of the study was to compare the effectiveness of four academic methods for elicitation and analysis of threats and security requirements - CORAS, SECURE TROPOS, SECURITY ARGUMENTATION and SI* - and to study their strengths and limitations. The participants were divided into groups composed by students and practitioners and were asked to identify the security requirements of a health care collaborative network using one of the methods under evaluation. The aim of eRISE challenge is similar to the one of our experiment: assessing methods' effectiveness. However, in eRISE effectiveness is assessed using a qualitative approach through questionnaires, post-it notes and focus group interviews with the participants. In our experiment, instead, we adopted a quantitative approach by using the number of high quality threats and security requirements identified by the participants as metrics to evaluate methods' effectiveness.

Other proposals focused on the comparison of methods for eliciting functional requirements [8], [9], [12]. Martinez et al. [8] conducted a quasi-experiment with 26 master students to compare three different paradigms for software development, model-driven, model-based and code-centric. The participants used each of the methods to develop a social network application. The authors assessed participant's perceived usefulness, perceived ease of use, perceived compatibility and intention to adopt through a post-task questionnaire. They also collected information about advantages and disadvantages of the methods reported by participants in the questionnaire. In our experiment, we have also measured perception variables (perceived easy of use, perceived usefulness and intention to use) with a post-task questionnaire. In addition, we assessed methods' effectiveness in terms of number of threats and security requirements identified by the participants.

Morandini et al. [9] carried out a controlled experiment to compare the effectiveness of Tropos and Tropos4AS in supporting the comprehension of requirements specifications. The experiment involved 12 participants including researchers and PhD students. The participants were asked to answer questions about comprehension of two requirements specifications drawn in Tropos and Tropos4S. In our study, we assessed methods' effectiveness in terms of number and quality of the threats and security requirements produced by the participants using visual and textual methods. It would be interesting to study how comprehensibility of requirements models impacts on the elicitation and analysis of threats and security requirements.

España et al. [12] conducted a laboratory experiment with 36 master students to compare two requirements engineering methods, Use Cases and Communication Analysis. Following the Method Evaluation Model (MEM) [13], the authors assessed methods' actual effectiveness and perceived efficacy. The actual effectiveness was evaluated by measuring the level of granularity and functional completeness of the functional requirements specification obtained with the two methods. In our experiment, instead, we evaluated the effectiveness as number of quality threats and security requirements generated by the participants using visual and textual methods. Similarly to España et al., we assessed perceived easy of use and perceived usefulness with a post-task questionnaire.

III. Research Method

This section describes the design of the performed experiment, following the guidelines by Wohlin et al. [14].

A. Selection of methods

CORAS is a visual method which consists of three tightly integrated parts, namely, 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 use diagrams (see Figure 1(a)) to document the result of the execution of each step. The steps are based on the international standard ISO 31000 [15] for risk management: context establishment, risk analysis (that identifies assets, unwanted incidents, threats and vulnerabilities), and risk treatments.

The Security Requirements Engineering Process (SREP) is an asset-based and risk-driven method for the establishment of security requirements in the development of secure Information Systems. SREP supports a micro-process, consisting of nine steps: agree on definitions, identify critical assets, identify security objectives, identify threats and develop artifacts, risk assessment, elicit security requirements, categorize and prioritize security requirements, requirements inspection, and repository improvement. The result of the execution of each step of the process is represented using tables or natural language (see Figure 1(b)). SREP is compliant with international standards ISO/IEC 27002 [16] and ISO/IEC 15408 [17] within the scope of requirements engineering and security management.

For additional details about CORAS and SREP we refer the reader to [2, Chap. 3] and [1]. Note that, in the rest of the paper, we denote with "security requirements" both the concepts "treatments" in CORAS and "security requirements" in SREP because they have the same semantic: they are both defined as a means to reduce the risk level associated with a threat.

B. Research approach

The *goal* of the experiment was to evaluate and compare two types of risk-driven methods, namely, visual methods (CORAS) and textual methods (SREP) with respect to their *effectiveness* in identifying threats and security requirements, and the participants' *perception* of the two methods. Hence, visual and textual methods were the two treatments that we have considered in the experiment. We want to investigate the following research questions:

RQ1 *Is the effectiveness of the methods significantly different between the two type of methods?*
RQ2 *Does the effectiveness of the methods vary with the assigned tasks?*
RQ3 *Is the participants' preference of the method significantly different between the two type of methods?*
RQ4 *Is the participants' perceived ease of use of the method significantly different between the two type of methods?*
RQ5 *Is the participants' perceived usefulness of the method significantly different between the two type of methods?*
RQ6 *Is the participants' intention to use the method significantly different between the two type of methods?*

To answer the first two research questions we have measured *effectiveness* by counting the number of threats and the number of security requirements as the main outcomes of the methods' application (as done in [7], [18]). Research questions $RQ3 - RQ6$ have been answered by measuring perception-based variables *perceived usefulness* (PU), *perceived ease of use* (PEOU), *intention to use* (ITU) with a post-task questionnaire. In order to gain a better understanding of *why a method is effective* (or more effective than another) we also carried out individual interviews with the participants.

C. Hypotheses

We have translated research questions $RQ1 - RQ6$ into a list of null hypotheses to be statistically tested. Due to the lack of space we report here only the main alternative hypotheses to the null ones denoted as Hn_A where n specifies the research question to which the hypothesis is related and the index A specifies that is an alternative hypothesis.

$H1.1_A$ There will be a difference in the number of threats found with the visual method and with the textual method
$H1.2_A$ There will be a difference in the number of security requirements found with the visual method and with the textual method
$H2.1_A$ There will be a difference in the number of threats found with the visual and the textual method within each facet
$H2.2_A$ There will be a difference in the number of security requirements found with the visual and the textual method within each facet
$H3_A$ There will be a difference in the participants preference for the visual and the textual method
$H4_A$ There will be a difference in the participants perceived ease of use for the visual and the textual method
$H5_A$ There will be a difference in the participants perceived usefulness for the visual and the textual method
$H6_A$ There will be a difference in the participants intention to use for the visual and the textual method

Hypotheses $H1.1_A$-$H1.2_A$ are related to $RQ1$ and suppose that there will be a difference in the effectiveness of the methods. $H2.1_A$-$H2.2_A$ assume a possible relation between the effectiveness of the methods and the facets on which the methods is applied (RQ2). Hypothesis $H3_A$ assumes there will be a difference in the participants' overall preference for the methods (RQ3). $H4_A$-$H6_A$ assume that the participants' perceived easy of use, perceived usefulness, and intention to use variables will differ for the two methods (RQ4-RQ6).

D. Experimental design

Participants for the experiments were recruited among master students enrolled in the Security Engineering course at the University of Trento. The participants had no previous knowledge of the methods under evaluation. A within-subject design where all participants apply both methods was chosen to ensure a sufficient number of observations to produce significant conclusions. In order to avoid learning effects, the participants had to identify threats and mitigations for different types of security facets of a Smart Grid application scenario. The Smart Grid is an electricity network that can integrate in a cost-efficient manner the behavior and actions of all users connected to it like generators, and consumers. They use information and communication technologies to optimize the transmission and distribution of electricity from suppliers to consumers.

The tasks differ in the security facets for which the groups had to identify threats and security requirements. The security facets included Security Management (Mgmnt), Application/Database Security (App/DB), Network/Telecommunication Security (Net/Teleco), and Mobile Security (Mobile). For example, in the App/DB facet, groups had to identify application and database security threats like

Facet/Method	Visual	Textual
Mgmnt	6	10
App/DB	9	7
Net/Teleco	9	7
Mobile	8	8

TABLE I: Experimental design

cross-site scripting or aggregation attacks and propose mitigations.

The participants were divided into 16 groups so that each group would apply the visual method (CORAS) to exactly two facets and the textual method (SREP) to the remaining two facets. For each facet, the method to be applied by the groups was randomly determined. Table I shows for each facet the number of groups assigned to visual and textual methods.

E. Experimental Procedure

The experiment was performed during the Security Engineering course held at University of Trento from September 2012 to January 2013. The experiment was organized in three main phases:

- **Training**. Participants were given a tutorial on the Smart Grid application scenario and a tutorial on visual and textual methods of the duration of two hours each. The Smart Grid scenario focused on the gathering of metering information from the smart meters and their transmission to the utility services for billing purposes. Then, participants were administered a questionnaire to collect information about their background and their previous knowledge of other methods and they were divided into groups based on the experimental design.

- **Application**. Once trained on the Smart Grid scenario and the methods, the groups had to repeat the application of the methods on four different facets: Security Management, Application/Database Security, Network Security and Mobile Security. For each facet, the groups:
  - Attended a two hours lecture on the threats and possible mitigations specific for the facet but not concretely applied to the case study.
  - Had one week to apply the assigned method to identify threats and security requirements 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 2013, each group submitted a final report documenting the application of the methods on the four facets.

- **Evaluation**. In this phase, the experimenters (the authors of this paper) assessed participants final reports while the participants evaluated the method through questionnaires and interviews. First, each group gave a presentation summarizing their work in front of the experimenters and of the expert. The expert evaluated the quality of the threats and the mitigations proposed for the Smart Grid application scenario. Then, participants were administered the post-task questionnaire to be filled in online. Last, 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, their advantages and disadvantages, the difficulties encountered during the application of the methods and the main differences among them. The interview questions were the same for all the interviewees even though some specific questions were added for some of the participants when their answers to the questionnaire were contradictory. The questions are reported in Table V in Appendix.

The questionnaire was adapted from the questionnaire reported in [7] which was inspired to the Technology Acceptance Model (TAM) [19]. The questionnaire consisted of 22 questions which were formulated in an opposite statements (positive statement on the right and negative statement on the left) format with answers on a 5-point Likert scale. The questions were formulated as follows: Q1: Whether the method was easy or hard to use; Q2: The method made the security analysis easier or harder than an ad hoc approach; Q3: The method was easy or difficult to master; Q4: Intention to use the method to identify threats and security requirements in a future project course; Q5: The method is better in identifying threats and security requirements than using common sense; Q6: Intention to use the method to identify threats and security requirements in a future project at work; Q7: Confusion about how to apply the method to the problem; Q8: Whether the method made the search for threats and security requirements more or less systematic; Q9: Intention to use the method if suggested by someone at work; Q10: Whether the method would be easy or hard to remember; Q11: Whether the method makes more or less productive in identifying threats and security requirements; Q12: Intention to use the method in a discussion with a customer; Q13: Whether the process of the method is well or not well detailed; Q14-Q15: A catalog of threats and security requirements makes easier or harder the security analysis with the method; Q16-Q17: The method helps or not helps in brainstorming on the threats and the security requirements; Q18: Whether the tool is easy or hard to use (asked just for the visual method because it had tool support); Q19-Q22: Difficulties of facets. To avoid that the participants answered on "auto-pilot", some of the questions (e.g. Q2, Q10, Q13) were given with the most positive response on the left and the most negative on the right.

IV. Reports' Analysis

In this section we report the results on methods' effectiveness based on the coding of groups' reports.

A. Coding

To assess the effectiveness of visual and textual methods, the final reports delivered by the groups were coded by the authors of this paper to count the number of threats and security requirements. An expert on security of the Smart Grid was asked to assess the quality of the threats and security requirements. The level of quality was evaluated on a four item scale: *Unclear* (1), *Generic* (2), *Specific* (3) and *Valuable* (4).

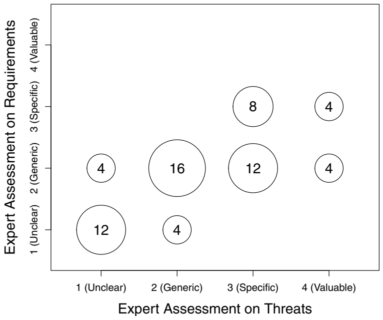


Fig. 2: Expert assessment.

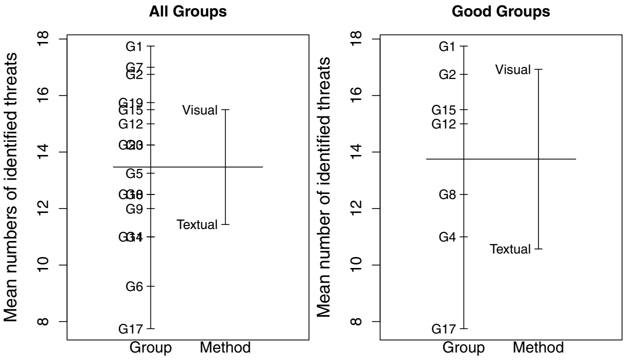


Fig. 3: Means of identified threats in all groups (left) and good groups (right).

Based on this scale, the groups who have got an assessment *Valuable* or *Specific* were classified as *good groups* because they have produced threats and security requirements of good quality. On the contrary, the groups who were assessed *Generic* or *Unclear* were considered as not so good (bad) groups.

Fig. 2 reports the expert assessment of all groups for all facets. In total we had 64 method applications because each of the 16 groups has applied one of the methods on the four facets. The number inside each bubble denotes the number of method applications which got a given expert's assessment for threats (reported on x-axis) and security requirements (reported on y-axis). There were 48 (75%) method applications that generated some clear threats (meaning threats evaluated generic, specific and valuable by the expert) while 28 (44%) method applications were specific to the scenario and appreciated by the expert. In contrast, the quality of produced security requirements was slightly lower than for threats: 48 (75%) method applications produced clear security requirements but more than half (36) were generic. In general, we can conclude that the overall quality of the outcomes of method applications was satisfactory.

B. Number of threats and security requirements

To test the effectiveness of visual and textual methods with respect to the number of identified threats and security requirements, we applied the ANOVA statistical test with a significance level α of 0.05. The ANOVA tables are not reported due to lack of space. Before the application of the test, we verified whether the dependent variables were normally distributed with Shapiro-Wilk test (returned *p-values* are 0.17, 0.68 for requirements and threats respectively). We also checked the homogeneity of variance with the Fligner-Killeen test (returned *p-values* are 0.45 for requirements, and 0.64 for threats). So we have no evidence to reject either assumptions.

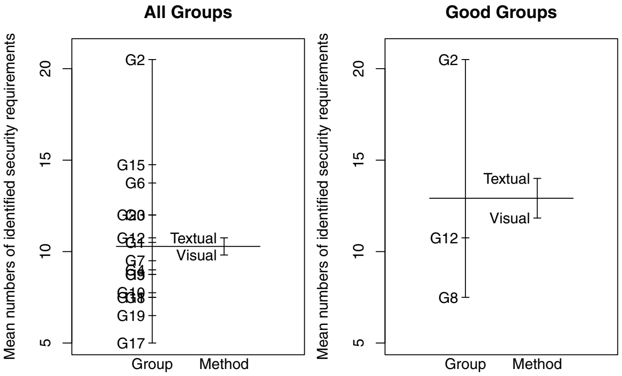


Fig. 4: Means of identified security requirements in all groups (left) and good groups (right).

We first analyzed the differences in the number of threats identified with visual and textual methods. As shown in Fig. 3 (left), if we consider all groups, the visual method is more effective in identifying threats than the textual one. This result is also confirmed if we consider only the groups who have produced good quality threats as shown in Fig. 3 (right). The ANOVA test shows that the effect of the applied methods on the number of identified threats is statistically significant for all groups ($F = 18.49,$ $p\text{-}value = 1.03 \cdot 10^{-4}$) and good groups ($F = 26.10,$ $p\text{-}value = 1.59 \cdot 10^{-4}$).

Similarly, Fig. 4 represents the means of the number of security requirements identified with the visual and the textual method by all groups (left) and by good groups (right). The figure shows that both for all groups and for good groups, the textual method is slightly better than the visual method in identifying security requirements. However, the ANOVA test shows that the difference in the security requirements identified with the textual and visual method is not statistically significant for all groups ($F = 1.18,$ $p\text{-}value = 0.28$) and good groups ($F = 1.98,$ $p\text{-}value = 0.23$).

Fig. 5 confirms the results shown in Fig. 3 and Fig. 4. The figure reports a scatter view of the distribution of identified security requirements, and identified threats for the groups which have applied the visual method (circles) and the one which have applied the textual method (triangles). The groups which applied the visual method tend to identify more threats, but less security requirements than groups which applied the textual method. The linear regression models on security requirements and threats show that the textual method is slightly better than the visual one in terms of the number of identified security requirements given the number of identified threats,

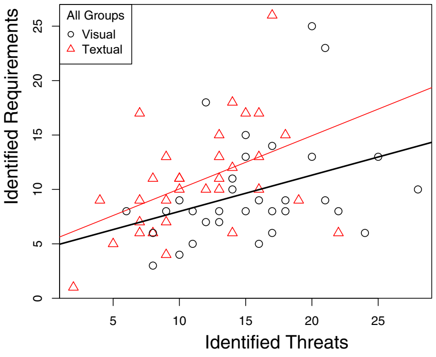


Fig. 5: Scatter plot of identified threats and security requirements for the two methods.

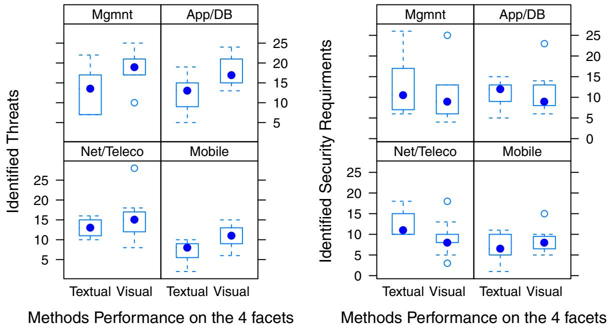


Fig. 6: The distribution of identified threats (left) and security requirements (right) within each facet.

but with no statistical significance.

We have also investigated the differences in the number of threats and security requirements identified with the visual and the textual method within each facet. The boxplots in Fig. 6 (left) show that the distribution of the visual method is always above the distribution of textual method. This means that using the visual method produces more threats than using the textual method in all four facets. This difference is less marked for the facet Net/Teleco (facet 3) but it is not statistically significant. If we consider only the facets Mgmnt, App/DB, and Mobile, the difference in the number of threats identified with the visual and the textual method is statistically significant because the ANOVA test returned a p-value $2.78 \cdot 10^{-3}$ ($F = 9.95$) which is less than 0.05. If we consider all facets, the difference is also statistically significant because the p-value returned by the ANOVA test is equal to $1.79 \cdot 10^{-3}$ ($F = 10.66$). Thus, we can conclude that across all facets the visual method is globally better than the textual method in identifying threats.

Fig. 6 (right) reports the number of security requirements identified with the visual and the textual method within each facet. The boxplots show that textual method is slightly better than the visual one in identifying security requirements in the first three facets. In particular, in facet Net/Teleco the difference is higher than in the facets Mgmnt and App/DB. However, the ANOVA test given the facet Net/Teleco returned $F = 3.37$, $\textit{p-value} = 0.09$, which means visual and textual methods distributions are different but the difference is not statistically significant. In the last facet Mobile, the situation is inverted and the visual method is better than the textual one in identifying security requirements. Given all facets, the ANOVA test returned $F = 0.57$, $\textit{p-value} = 0.45$ which means the difference in the number of security requirements is not statistically significant.

		All subjects			Good subjects		
Q	**Type**	**Mean**		**Z**	**Mean**		**Z**
		Textual	**Visual**		**Textual**	**Visual**	
1	PEOU	3.0	3.3	-0.9	2.8	3.8	-2.1 *
2	PU	3.1	3.6	-2.1 *	3.3	3.5	-0.6
3	PEOU	3.3	3.1	0.4	3.5	3.8	-0.7
4	ITU	3.0	3.1	-0.2	2.8	3.4	-1.5
5	PU	3.1	3.0	0.3	2.9	3.6	-2.3 *
6	ITU	2.9	2.9	0.0	2.5	3.0	-1.3
7	PEOU	3.0	3.1	-0.2	3.2	3.2	0.0
8	PU	3.6	3.5	0.4	3.8	3.8	0.0
9	ITU	3.2	3.4	-0.9	3.2	3.5	-1.3
10	PEOU	3.5	3.7	-0.5	3.8	3.9	0.2
11	PU	3.1	3.2	-0.4	2.8	3.2	-1.0
12	ITU	3.0	3.3	-0.8	2.9	3.3	-1.0
13	Control	3.8	3.8	-0.2	3.5	4.2	-1.7
14	Control	4.4	3.9	3.1 ***	4.5	3.8	2.3 *
15	Control	4.3	4.2	0.9	4.5	4.3	1.3
16	Control	3.0	3.6	-2.5 *	2.9	3.7	-1.8
17	Control	3.1	3.5	-1.7	3.2	3.8	-2.3 *
PEOU		3.2	3.3	-0.7	3.3	3.7	-1.7 •
PU		3.2	3.4	-1.0	3.2	3.5	-2.0 •
ITU		3.0	3.2	-1.0	2.8	3.3	-2.5 *
Total		3.2	3.3	-1.6 •	3.1	3.5	-3.5 ***

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

TABLE II: Wilcoxon signed-ranks test of responses

V. Questionnaire Analysis

We have analyzed the responses to the post-task questionnaire to determine if there is a difference in the participants' perception of visual and textual methods. When reporting the results all answers have been realigned to 5 being the best. As the responses were paired, in general not normally distributed, and our samples had ties, we have used the exact Wilcoxon signed-ranks test with Wilcoxon method for handling ties [20]. We set the significance level α to 0.05. As mentioned, for some of the questions (e.g. Q2 or Q10), we had to invert the order of negative and positive responses so that for all questions all negative responses were on the right and positive responses were on the left.

The results are summarized and compared in Table II. For each question, the table reports to which perception variable the question refers to (PEOU, PU, ITU), the mean of the answers by all and by good participants (the one who were part of groups that produced good quality threats and security requirements based on expert's assessment), and the level of statistical significance based on the p-value returned by the Wilcoxon test. The level of statistical significance is specified by • (p<0.1), or * (* p<0.05, ** p<0.01, *** p<0.001). The table also reports the average responses for each perception variable and for all questions related to perception (Q1-Q12).

The results show that for some aspects the difference in the perception of visual (CORAS) and textual method

(SREP) is statistically significant (p<0.05) or has minimum 10% significance level:

Q1 All participants prefer visual method over the textual one for easy of use but the difference in perception is not statistically significant. Instead, for good participants the difference is statistically significant.

Q2 Visual method is better than textual approach with respect to making the security analysis easier than an ad hoc approach. All participants prefer visual method with statistical significance. This is also true for good participants but the difference in participants' perception is not statistically significant.

Q5 When considering finding threats and security requirements more quickly than using common sense the results are not clear. The results show a small preference for textual method by all participants which is not statistically significant. In contrast, good participants show a statistically significant preference for visual method.

Q14 Both all participants and good participants with statistical significance believe that a catalog of threats would be more needed by textual method than the visual one.

Q16 Visual method is better than textual one with respect to helping brainstorming on the threats. This holds across all participants and the difference of preference is statistically significant. This is also true for good participants but with no statistical significance.

Q17 Visual method is better than the textual one with respect to helping brainstorming on the security requirements. This holds across all participants but it is not statistically significant. This is also true for good participants and it is statistically significant.

PEOU Visual method is better than the textual method with respect to overall PEOU across all participants but the preference is not statistically significant. Good participants show a small preference for visual method with 10% significance level.

PU Visual method is better than the textual one with respect to overall PU across all participants but the preference is not statistically significant. Good participants show a small preference for visual method with 10% significance level.

ITU Visual method is better than the textual one with respect to overall ITU across all participants but the preference is not statistically significant. Good participants show a statistically significant preference for visual method.

The average responses to Q1-Q12 show a small preference for visual method by all participants with 10% significance level. For good participants, instead the preference for visual method is statistically significant.

VI. Interviews' Analysis

For a better understanding of which features influence visual and textual methods effectiveness, we complemented our experiment by interviewing each participant for half an hour.

The interviews were analyzed with a content analysis technique called coding [21]. The analysis consists of the following steps: *1*) we transcribed and analyzed to identify recurring themes, which serve as the basis to build categories that explain why visual and textual methods work in practice or not; *2*) we identified a set of recurring participants' statements in the interviews and we classified them in *advantages*, *disadvantages* and *improvements* of the methods; *3*) for each group of statements, we coded and classified them into iteratively emerging categories; *4*) we counted the frequency of statements in each category as an indication of their relative importance. Table III presents the categories and the frequency of statements in each category made by the participants.

TABLE III: Frequency of reported aspects of the methods

Advantages	**Visual**	**Textual**	**Total**
Clear process	12	16	28
Help in brainstorming threats	21	15	36
Help in brainstorming security requirements	12	24	36
Easy to use and remember	17	11	28
Help to understand interdependencies	6	7	13
Support visual summary	24	0	24
No time consuming	0	4	4
Total	92	77	
Disadvantages			
No clear process	2	11	14
Do not support interdependencies	2	3	5
No help in brainstorming threats	3	3	6
No help in brainstorming security requirements	9	1	10
Primitive tool	20	0	20
No support visual summary	1	6	7
Visual summary does not scale	10	0	10
Too time consuming	11	9	20
No easy to use and remember	0	2	2
Total	58	35	
Improvements			
Have security resource repository	0	5	5
Have visual summary	0	2	2
Support automatic risk level computation	1	0	1
Support diagram creation	1	0	1
Total	2	7	

The main advantage of visual method that participants indicated is that it provides a visual summary of the results of the security analysis (89%). Indeed, the diagrams give an overview of the assets and the possible threats scenarios and treatments. A typical statement made by the participants referring to this advantage was: "*Diagrams are useful. You have an overview of the possible threat scenarios and you can find links among the scenarios*". Another noteworthy advantage of visual method reported by the 82% of the participants was that it helps brainstorming on the threats. As the participants indicated, diagrams play a key role in helping to brainstorm on threats: "*Yes it helped to identify which are the threats. In CORAS method everything is visualized. The diagrams helped brainstorming on threats*." The next advantage refers to perceived easy of use. The 60% of the participants reported that visual method is a "*good methodology, not difficult to use. It is much clear to understand the security case there*".

The main advantage of textual method according to the 96% participants was that the method helps in identifying security requirements. Typical statements in this category were: "*SREP helped in brainstorming. The steps were pretty much defined. Step by step helped to discover more*" and "*SREP helped in brainstorming. The order of the steps helped to identify security requirements*". The second advantage of textual method is that it has a clear process to follow (60%): "*Well defined steps. Clear process to follow.*" is an example of typical statement made by the participants for this category.

TABLE IV: Results of hypothesis testing

$H1.1_A$	Difference in the number of threats found with visual and with textual method	YES (More threats were found with visual method than with textual method)
$H1.2_A$	Difference in the number of security requirements found with visual and with textual method	NO (Slightly more security requirements were found with textual method than with the visual one but the difference is not statistically significant)
$H2.1_A$	Difference in the number of threats found with visual and with textual method within each facet	YES (For each facet more threats were found with visual than with textual method)
$H2.2_A$	Difference in the number of security requirements found with visual and with textual method within each facet	NO (For each facet slightly more security requirements were found with textual than with visual method but the difference is not statistically significant)
$H3_A$	Difference in the participants preference for visual and textual method	YES (Overall visual method is preferred to the textual one)
$H4_A$	Difference in the participants perceived ease of use for visual and textual method	MAY BE (Visual method is perceived as easier to use than textual approach with 10% significance level)
$H5_A$	Difference in the participants perceived usefulness for visual and textual method	MAY BE (Visual method is perceived as more useful than the textual method with 10% significance level)
$H6_A$	Difference in the participants intention to use for visual and textual method	YES (Participants intend to use the visual method more than the textual one)

With respect to methods' disadvantages and improvements, the statements were fewer than the ones about advantages. The most indicated disadvantage of visual method was that visual notation does not scale well for complex scenarios. Typical statements in this category were: "*The diagrams are not scalable when there are too many links*" and "*For big systems the diagrams would be very large. Even with the support of the computer it would be difficult to see them.* In addition, 75% of the participants complained about the tool. The major problems reported were the tool bad memory usage that makes the tool too slow and the modeling feature of the tool that does not provide automatic support for the generation of the diagrams (e.g. generating a treatment diagram from a threat diagram). Examples of typical statements for this category were: "*The tool is not difficult to use but it is very slow. It is impossible to copy a diagram from a type of diagram to another. Objects have no references between the diagrams. Changes on an object in a diagram are not reflected on the same object in other diagrams.*" and "*The tool takes too much to arrange things. Drawing assets and threats is not easy. When the diagrams are too large, the tool occupies too much memory*". Instead, textual method has two main drawbacks. First, it is unclear how to perform some of the steps of the textual method process: risk assessment, requirements inspection and repository improvement. Second, the use of tables to represent threats makes it difficult to show the link among assets, threats and security requirements, and thus to give a summary of the results of the security analysis. As reported by the participants "*It is not easy to represent what you think because there are a lot of tables. If you are project manager and you want to show the results of the security analysis to your boss it is difficult because you use tables*".

VII. Discussion

In this section we present the main findings regarding each of the research questions and possible explanations for the findings. A summary of the findings is shown in Table IV.

A. Methods' effectiveness

As shown in the previous sections, visual method is more effective in identifying threats than textual method. This result is also confirmed if we consider the *number of threats* identified with visual and textual methods across the task assigned to the groups. Since the difference in the number of threats identified with the two methods is statistically significant, we can accept the alternative hypotheses $H1.1_A$ and $H2.1_A$ of difference between the number of threats identified with the two methods. Instead, with respect to *number of security requirements*, textual method is slightly more effective than the visual one in identifying security requirements but the difference is not statistically significant across all groups and tasks. The alternative hypotheses $H1.2_A$ and $H2.2_A$ of difference in the number of security requirements can therefore be rejected.

B. Methods' perception

Participants' *overall preference* is higher for visual than for textual method. Among all the groups the difference has 10% significance level, while for the participants who were part of groups who produced good quality threats and security requirements, the difference in the overall preference is statistically significant. The conclusion is that the alternative hypothesis $H3_A$ of difference in the overall preference of the two methods is upheld. Similarly, for all participants there is no statistically significant difference in perceived *easy of use* and *usefulness*, while for "good" participants the difference has a 10% significance level. For this reason, there is no evidence that the null hypotheses $H4_0$ of no difference in the perceived easy of use and $H5_0$ of no difference in perceived usefulness do not hold. Thus, the alternative hypotheses $H4_A$ and $H5_A$ cannot be rejected or accepted. With respect to *intention to use*, "good" participants intend to use more visual than textual method and the difference in participants' perception is statistically significant. The alternative hypothesis $H6_A$ of difference in the intention to use for the two methods can thus be accepted.

C. Qualitative Explanation

The different number of threats and security requirements identified with visual and textual methods can be likely explained by the differences between the two methods indicated by the participants during the interviews. Diagrams in visual method help brainstorming on the threats 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. As suggested by the answers to question $Q14$ in the post-task questionnaire, the identification of threats in textual method could be made easier if a catalog of common threats was available. In addition, during the interviews some of the participants indicated that a visual representation for threats would be better than a tabular one.

Textual method is slightly more effective in eliciting security requirements than visual approach because the order of steps in textual method process guides the analyst in the identification of security requirements, while the same it seems not to hold for the visual method's process.

VIII. Threats to Validity

We discuss the four main types of threats to validity [14] in what follows.

a) Conclusion validity: Conclusion validity is concerned with issues that affect the ability to draw the correct conclusion about the relations between the treatment and the outcome of the experiment. There are three main threats to conclusion validity relevant for our experiment:

- *Low statistical power*. An important threat to validity is related to the sample size that must be big enough to come to correct conclusions. We conducted a post-hoc power analysis for the ANOVA test and Wilcoxon signed-rank test (with G*Power 3 tool[1]) for participants from good groups. For Wilcoxon signed-rank test, we obtained a power (1-β) equal to 0.86 setting as parameter the effect size ES = 0.71, the total sample size N = 24, and α = 0.05. For the ANOVA test, we have instead a power of 0.89 with 32 observations for each method and between variance at least 16 observations are needed to have an effect size of 2 like in our experiment. We thus have enough observations to conclude that our results on the relation between the methods applied and their performance in terms of number of threats and security requirements and with responses to the post-task questionnaire are correct.
- *Violated assumptions of statistical tests*. Before running the ANOVA and Wilcoxon signed rank tests we have checked with Shapiro-Wilk and Flinger-Killeen tests that their assumptions are not violated.
- *Heterogeneity of subjects*. If groups in the sample are too heterogeneous, the variation due to individual differences may be larger than due to treatment. We have reduced this threat by running the experiment with master students who had similar knowledge and background.

b) Internal validity: Internal validity is concerned with issues that may falsely indicate a causal relationship between the treatment and the outcome, although there is none.

- *Participants' background*. The familiarity of the participants with the methods evaluated during the experiment is a threat to internal validity. At the beginning of the the experiment, we have administered a questionnaire to check the background of the participants and their knowledge of security methods. The questionnaire has shown that all participants had a similar background and had no prior knowledge about visual and textual methods.
- *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.
- *Participants' behavior*. During the execution of the experiment, the subjects may react differently over time e.g., subjects may become bored or tired, or they may become more or less positive to one or another method. We notice that the performance of the participants in terms of number of threats and security requirements identified was almost the same for the first, second and third task, while on the last task the performance decreases because the participants got tired or did not put much effort. Yet, this phenomenon was common to both methods.
- *Bias in data analysis*. To avoid bias in the reports analysis, the coding of the participants' reports was conducted by the authors of the paper independently. In addition, the quality of the threats and security requirements identified by each group was assessed by an expert external to the experiment.

c) Construct validity: Construct validity concerns generalizing the result of the experiment to the concept and theory behind the experiment. The main threat to construct validity in our experiment is the design of the research instruments: interviews and questionnaires. The questionnaire was designed following the Technology Acceptance Model with four questions for each of the independent variables we wanted to measure: *perceived usefulness*, *perceived easy of use*, *intention to use*. The interview guide included questions concerning research questions $RQ3$ and methods' advantages and disadvantages. Three researchers independently have 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.

d) External validity: External validity concerns the ability to generalize experiment results beyond the experiment settings. External validity is thus affected by the objects and the subjects chosen to conduct the experiment.

- *Use of students instead of practitioners*. Using students rather than practitioners as subjects is known as a major threat to external validity. However, Svahnberg et al. [22] recognized that students may work well as subjects in empirical studies in the requirements engineering area.
- *Realism of the application scenario and facets*. We reduce the threat to external validity by making the experimental environment as realistic as possible. In fact, as object of our experiment we have chosen a real industrial application scenario proposed by National Grid. Furthermore, the reports of participants have been evaluated by an expert from National Grid: the quality of both security requirements and threats identified is good enough for the study (see also §IV).

IX. Conclusions

We conducted a controlled experiment with 28 master students in computer science to investigate the *effectiveness* and

[1] http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

the participants' *perception* of two type of risk-based methods, visual methods (CORAS) and textual methods (SREP). The participants were divided into 16 groups and had to solve four different tasks that required the identification of threats and security requirements for different security facets of a real Smart Grid application scenario. The number of threats and security requirements identified by each group was used to evaluate the effectiveness of the two methods. Participants' perception was assessed through a post-task questionnaire. In addition, each participant was interviewed to gain insights about why the methods are effective, or why they are not.

The main result was that the visual method is more effective than the textual method in threat identification. From the analysis of participants' interviews, the visual method is more effective in identifying threats because threats diagrams help brainstorming on threats. The textual method, in contrast, is slightly more effective in identifying security requirements than the visual method because the process guides the analyst to the identification of good quality security requirements. The visual method is also overall preferred to the textual method as shown by the analysis of the post-task questionnaire.

We are planning to analyze in depth the *whys* of our results. To do this, we will carry out other evaluations where we assess the methods' effectiveness and we analyze the relations between effectiveness and methods' aspects indicated by the participants as advantages and disadvantages of the methods.

Acknowledgment

This work is funded by EU-FP7-IST-NoE-NESSOS, and EU-FP7-SEC-SECONOMICS. We would like to thank Dr. Raminder Ruprai from National Grid for participating as an expert.

References

[1] D. Mellado, E. Fernández-Medina, and M. Piattini, "Applying a security requirements engineering process," in *Proc. of the 11th European Symposium on Research in Computer Security (ESORICS)*. Springer, 2006, pp. 192–206.

[2] M. S. Lund, B. Solhaug, and K. Stølen, *Model-driven risk analysis: the CORAS approach*. Springer, 2011.

[3] H. Mouratidis, "Secure software systems engineering: The secure tropos approach," *Journal of Software*, vol. 6, no. 3, pp. 331–339, 2011.

[4] P. Giorgini, F. Massacci, J. Mylopoulos, and N. Zannone, "Modeling security requirements through ownership, permission and delegation," in *In Proc. of the 13th IEEE International Conference on RE*. IEEE, 2005, pp. 167–176.

[5] M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen, "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements," *Requirements Engineering*, vol. 16, no. 1, pp. 3–32, 2011.

[6] G. Sindre and A. Opdahl, "Eliciting security requirements with misuse cases," *Requirements Engineering*, vol. 10, no. 1, pp. 34–44, 2005.

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

[8] Y. Martínez, C. Cachero, and S. Meliá, "Mdd vs. traditional software development: A practitioner's subjective perspective," *Information and Software Technology*, 2012.

[9] M. Morandini, A. Marchetto, and A. Perini, "Requirements comprehension: A controlled experiment on conceptual modeling methods," in *In Proc. of the International Workshop on Empirical RE*. IEEE, 2011, pp. 53–60.

[10] F. Massacci and F. Paci, "How to select a security requirements method? a comparative study with students and practitioners," in *Secure IT Systems*. Springer, 2012, pp. 89–104.

[11] N. Condori-Fernandez, M. Daneva, K. Sikkel, R. Wieringa, O. Dieste, , and O. Pastor, "A systematic mapping study on empirical evaluation of software requirements specifications techniques," in *In Proc. of the Third International Symposium on ESEM*. ACM, IEEE, 2009, pp. 502–505.

[12] S. España, N. Condori-Fernandez, A. González, and Ó. Pastor, "An empirical comparative evaluation of requirements engineering methods," *Journal of the Brazilian Computer Society*, vol. 16, no. 1, pp. 3–19, 2010.

[13] D. L. Moody, "The method evaluation model: a theoretical model for validating information systems design methods," in *In Proc. of the 11th European Conference on IS (ECIS)*, 2003, pp. 1327–1336.

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

[15] *ISO 31000 Risk management – Principles and guidelines*, International Organization for Standardization, 2009. [Online]. Available: http://en.wikipedia.org/wiki/ISO_31000

[16] *ISO/IEC 27002 Information technology - Security techniques - Code of practice for information security management*, International Organization for Standardization and International Electrotechnical Commission, 2005. [Online]. Available: http://en.wikipedia.org/wiki/ISO/IEC_27002

[17] *ISO/IEC 15408 Information technology Security techniques Evaluation criteria for IT security*, International Organization for Standardization and International Electrotechnical Commission, 2005. [Online]. Available: http://www.enisa.europa.eu/activities/risk-management/current-risk/laws-regulation/rm-ra-standards/iso-iec-standard-15408

[18] A. Teh, E. Baniassad, D. Van Rooy, and C. Boughton, "Social psychology and software teams: Establishing task-effective group norms," *Software, IEEE*, vol. 29, no. 4, pp. 53–58, 2012.

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

[20] W. J. Conover, "On methods of handling ties in the wilcoxon signed-rank test," *Journal of the American Statistical Association*, vol. 68, no. 344, pp. 985–988, 1973.

[21] A. L. Strauss and J. M. Corbin, *Basics of qualitative research: techniques and procedures for developing grounded theory*. Sage Publications, Thousand Oaks, Calif, 1998.

[22] M. Svahnberg, A. Aurum, and C. Wohlin, "Using students as subjects-an empirical evaluation," in *Proceedings of the Second International Symposium on ESEM*. ACM, IEEE, 2008, pp. 288–290.

Appendix

Interview Questions
What do you think about method?
Do you think the method is an easy method to apply? Why?
While applying the method where you got confused about how to apply it?
Do you think the method helps you brainstorming? Why?
Do you think the method helped you to identify threats and security requirements?
Which are the advantages of the method?
Which are the disadvantages of the method?
Would you use the method in the future?
What do you think about CORAS tool?
Do you think CORAS tool is hard to use? Why?
Which version of the CORAS tool did you use?
Which do you think are the significant differences between the two methods?
Which was according to you the most difficult facet? And why?

Note: These questions were asked both for the visual (CORAS) and the textual method (SREP).

TABLE V: Interview Guide