Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CATALOGUE-BASED GUIDELINE FOR MISUSE CASE

[1] MUHAMMAD ASYRAF KHAIRUDDIN, [2] ABDUL AZIM ABD GHANI, [2] HAZURA ZULZALIL, [2] SA'ADAH HASSAN

[1]Software Engineering Programme, Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak, Malaysia
[2]Software Engineering and Information System, Faculty of Computer Science & Information Technology, Universiti Putra Malaysia, Malaysia
E-mail: [1]kmasyraf@unimas.my, [2]{azim, hazura, saadah}@upm.edu.my

ABSTRACT

Misuse case is one of the security requirement elicitation techniques that are easy to use and learn. Unfortunately, the current guideline provided is too general. The process of identifying the misuse case and threats is open for the analyst's interpretation. Lack of knowledge in security threats also can make it worse. These problems can lead to analysis paralysis situation. In this paper, we proposed a catalogue-based guideline to support misuse case techniques to elicit security requirements. This guideline consists of two catalogues used to assist software developers in identifying attacks and threats from a misuse case diagram. We experimented with selected students to evaluate the effectiveness of the guideline in identifying threats and types of threats. We also evaluated the usability of the guideline by conducting experts reviews. Experiment's result shows sufficient evidence that using the misuse case with the proposed catalogue-based guideline is more effective in identifying threats and types of threats than using the misuse case without a guideline. Expert review's result also shows that the catalogue-based guideline is more usable in identifying threats than without using the guideline.

Keywords: *Misuse Case, Security Requirements, Threats, Catalogue, Guideline*

1. INTRODUCTION

One of the ways to elicit security requirements is by using the misuse case technique. Misuse case is easy to learn and understand by software developers as it is based on the commonly use Unified Modeling Language (UML) use case diagram. It can be considered as an extension to a use case diagram. By using a misuse case diagram, we only need to have several brainstorming sessions before we can identify threats or attacks that can happen to the system. However, there are several problems in which can hamper its performance to identify threats. The first problem is there are no exact guidelines provided to elicit security requirements. The instruction is too general and imprecise [1], where the process of identifying misuse cases and threats are open for the analyst' interpretation. This problem may lead to either the development of insecure software or analysis paralysis. Analysis paralysis often occurs due to overanalyzing or overthinking for a solution but end up with too many solutions and cannot decide which solution to take on. Thus, a proper guideline is needed to avoid analysis paralysis from happening.

The second problem is the knowledge gap in the security field. In order to identify the threats, the analyst needs to have particular knowledge of threats and attacks. That means the identifying process is depended on the analyst experience and knowledge. Most of the software developers do not have sufficient knowledge of security. Since more reliable and secure software is needed, software developers need to equip their development team with a security expert. However, getting the security expert to join the team may need extra cost for a smaller development team.

This paper aims to evaluate the catalogue-based guideline for misuse cases in identifying threats and types of threats and whether the proposed guideline is useful in identifying the threats.

2. RELATED WORKS

Several authors have proposed solutions and enhancements to improve the use of the misuse

case technique. To solve a problem regarding inexperienced developers in security threats, [2] have proposed a method using Common Criteria and related knowledge sources to identify security requirements from functional requirements through eliciting threats and security objectives. However, there are several weaknesses in this technique. First, the attributes and the inference rules need to be explored and elaborated. Second, the application of the ontological approaches needs to be investigated. Third, the other types of knowledge sources for security requirements to be integrated with the technique need to be considered. And lastly, more practical case studies need to be done to evaluate the technique.

[3], [4] proposed a new "Hybrid Technique" to improve misuse case usability in a large system. This Hybrid Technique merges misuse cases and attack trees' strengths, strengthening the system to mitigate weaknesses effectively in large and complex systems. Nevertheless, this technique still did not have any tool at that time, and no evaluation has been done.

Another group of researchers [5] proposed a way to bridge the gap between software developers and security experts by providing security knowledge in the form of reusable security models and tool artefacts aimed at different software development phases. Their basic ideas are that security experts produce and link threat models, i.e., security knowledge, documented via easily understandable, informative models, which can be reused by software developers and development teams to acquire the information they need. [5] addressed how existing threat models, i.e., misuse cases and attack trees, can be used together and linked to show high-level and more detailed threats towards standard software functionality. They also show how threats can be linked to UML activity diagrams to model development activities for threat mitigation purposes. The only problem with this technique is that it is still not evaluated yet and needs more experiments for that purpose.

Lack of security expertise, security requirements are too vague or overly specific, or even neglected are challenges that motivate [6] to propose tool-based support for the security requirements specification (SRS) that could facilitate the generation of proper quality security requirements, and reduce the amount of effort. They present an ontology-based approach that uses predefined pattern-based templates – requirements boilerplates – to aid requirements engineers in the formulation of SR. They applied the technique to a prototype tool that enables the formulation of security requirements from textual misuse case (TMUC) descriptions of security threat scenarios. The evaluation that has been done is to assess how well the approach complements the requirement analyst by reducing the effort needed for SRS and stimulating the specification of formal quality security requirements. It got a positive result and only received minor comments regarding the tool platform's limitation, annotation of the terms, and helpful tips. More improvement on the ontology side is also needed.

Unable to prioritize security requirements motivate [7] to propose an enhanced misuse case that can integrate a way to prioritize security requirements based on the budget availability. This solution can help software developers prioritize which requirements need to develop first. In order to elicit security requirements effectively, [8] proposed a different approach where they integrate misuse cases and attack patterns with threat modelling. They also investigate how misuse cases can enhance the performance of threat modelling. Another researcher, [9], try to construct security test models from the artefact of misuse case modelling. The security test model then can automatically generate security tests where the test inputs are from misuse cases.

As misuse case models are prone to human error,[10] proposed a model transformation to apply changes in misuse case models if it got design issues. The solution is to detect antipatterns and apply refactorings to eliminate the errors. This model will increase the quality of the misuse case diagram to elicit security requirements by the developers.

Most of the works do not improve the misuse case but propose new ways to implement the security requirement elicitation technique. [2],[4],[5] and [8] try to increase the misuse case ability to detect threats. However, they did not discuss how effective and how large the size of the coverage of their technique was. [6] proposed tool-based support to generate quality security requirement specifications. [7],[9] and [10] enhance misuse cases' ability to prioritize requirements, construct security tests, and detect design errors.

3. RESEARCH METHODOLOGY

The research activities involved in this work are divided into three stages (refer to Figure 1).

3.1 Literature Review

The initial stage to start the research is to conduct a literature review. This stage is to observe and get enough knowledge on security requirement elicitation techniques, especially the misuse case, and the other supporting techniques from existing research that can improve the misuse case. Brief introductions on requirement elicitation, software security, and security threats are also included to understand the background of the research problems. All the findings from the literature review were analyzed and were taken into account during guideline design.

3.2 Design And Development

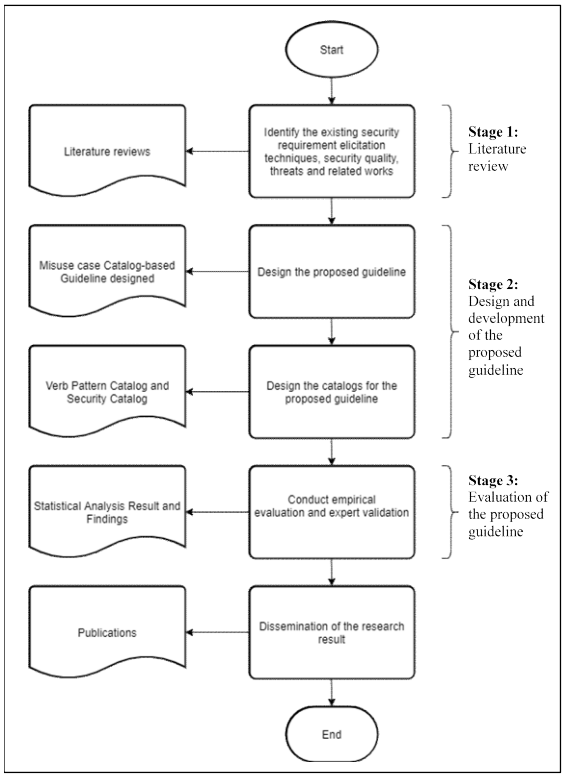

Figure 1: Research activities

The second stage is to design and develop the guideline. Based on the findings from the literature review, a guideline framework called catalogue-based guideline has been proposed. Before that, we also need to know what type of security attribute that we need to secure in the developed system. For this research, we choose the CIA triads (Confidentiality, Integrity and Availability) as it is considered as the essential element of security controls [11]. Table 1 shows the definition for the chosen security attributes. These security attributes will be used in two catalogues; the Verb Pattern catalogue and the Security catalogue.

The purpose of using catalogues in the proposed guideline is to equip the user with knowledge on security during the process of finding the misuse case. The word-matching technique is to create the Verb Pattern catalogue and the attack tree technique is to create the Security catalogue.

3.2.1 Word-matching technique

The word-matching technique is a technique that uses a selected word (verb) and matches it to a specific partner (security attribute). These matching created a pattern, hence the name Verb Pattern catalogue. This technique is inspired by [2] and [13]. [2] used word-matching to match the word captured in a use case description with the word in the catalogue. Meanwhile, [13] used the technique to maps important key phrases to appropriate essential requirements for their Essential Interaction Library in their work.

Table 1: Definition of each security [12].

Security Attributes	Definition
Confidentiality	the property that information is not made available or disclosed to unauthorized individuals, entities, or processes.
Integrity	the property of safeguarding the accuracy and completeness of assets.
Availability	the property of being accessible and usable upon demand by an authorized entity.

In order to create the catalogue, 250 use cases from 3 domains, which are healthcare, business, and e-commerce system, have been collected. From these use cases' names, the verbs and nouns that were used in the name were identified. Once the verbs are already identified, we need to do a brainstorming session where we need to match the verbs with the chosen security attributes. To match up, we need to think about what security attributes are involved if the verb is paired with a noun similar to the identified noun. To do this activity, we need to use the shared knowledge and also our experiences in dealing with

security threats. The results of this activity are the list of match up verbs with security attributes, also known as the Verb Pattern catalogue (refer to Table 2).

Table 2: Example of Verb Pattern catalogue.

No	Verb	Security Attributes
1	Add	Confidentiality, Integrity
2	Admit	Confidentiality, Integrity
3	Apply	Availability
4	Assign	Confidentiality, Integrity
5	Authenticate	Integrity
6	Backup	Confidentiality, Integrity
7	Bill	Integrity
8	Browse	Availability
9	Calculate	Integrity
:	:	:

3.2.2 Attack tree

The attack tree is a diagram that consists of one root node or parent node as the goal and ways to achieve the goal as leaf nodes or child nodes [14]. Figure 2 shows the example of the attack trees in identifying threats. Parent nodes ("Flood system" and "Steal info") show what the attack is, and the child nodes show how it attacks.

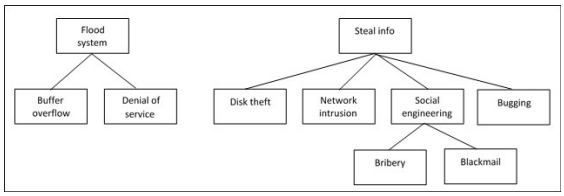

Figure 2: Example of attack trees

In this research, the attack tree is used to assist in the creation of the Security catalogue. The parent node is defined as types of attack, and the child nodes are defined as the attack or threat that can happen. In order to create the Security catalogue, 11 types of attacks and 81 attacks and threats have been identified from a total of 7 security information websites (refer to Table 3 for the definition of types of threats or attacks used in this research).

The types of threats or attacks, which also will be used as misuse cases in the guideline later, is used as a root for an attack tree in order to find the threats that can happen through the misuse cases. The Security catalogue also contains the chosen security attributes (CIA triads) that were mapped to the types of threats (parent node), which was done through a brainstorming session.

Table 3: Definition for each types of threat or attack

No.	Types of Threat or Attack	Definition
1.	Bluetooth related attacks	Any threats or attacks that are made through Bluetooth technology.
2.	Cyber fraud	A crime committed through a computer with the intent to corrupt another individual's personal and financial information.
3.	Flooding attacks	Overwhelm the victim's network to disrupt the service.
4.	Login attack	Any threats or attacks that are made towards the login page to access the system.
5.	Malicious content	Materials that are not suitable for an average reader include hate speech, violence, porn, et cetera.
6.	Malware attacks and infections	Any type of malicious software used to get information, breach privacy, or disrupt service.
7.	Password attack	Any threats or attacks that are made towards the password field to access the system.
8.	Vulnerability exploitation	To exploit or takes advantage of a software vulnerability or security flaw.
9.	Physical attacks	Any threats or attacks that are done physically.
10.	Social engineering attacks	Use deception to manipulate individuals to get personal or confidential information.
11.	Technical attack	An attack can be made by circumventing or nullifying hardware and software protection mechanisms rather than by subverting system personnel or other users.

3.3 Evaluation

Proposing a guideline must be accompanied by an analysis of the guideline's effectiveness and experts' reviews before it can be considered for adoption. This section describes an experiment conducted to evaluate the guideline and the evaluation by experts on the usability of the proposed guideline.

3.3.1 Empirical evaluation

This subsection discussed the experiment definition, planning, execution, and validity threats to the experiment.

3.3.1.1 Experimental Definition

This experiment's goal is to compare the misuse case technique catalogue-based guideline for its effectiveness in identifying threats and its

coverage. These evaluation criteria are nearly similar to what [15] did. In this study, effectiveness refers to the capability of using the technique to identify threats per execution time through the proposed guideline, and coverage refers to the capability of using the technique to identify types of threat per execution time through the proposed guideline. This investigation is essential to find out whether using the proposed misuse case technique with the catalogue-based guideline can deliver better results than without using the guideline. To achieve the stated goal, we set to investigate the following research questions:

- Does using the catalogue-based guideline increase the ability to identify threats compared to without using the guideline?
- Does using the catalogue-based guideline increase the ability to identify types of threats compared to without using a guideline?

In order to address RQ1 and RQ2, the following hypotheses can be determined:

- *Null hypothesis* (H_{0Eff}) = There is no significant difference in the effectiveness of identifying threats using the proposed misuse case catalogue-based guideline and without using the proposed guideline. It can be formulated as H_{0Eff}: $\mu_{diff} = 0$, where μ_{diff} is the mean of the difference in the number of threats identified by each subject.
- *Alternative hypothesis* (H_{1Eff}) = There is a significant difference between the effectiveness of identifying threats using the proposed misuse case catalogue-based guideline and without using the guideline. It can be formulated as H_{1Eff}: $\mu_{diff} > 0$, where μ_{diff} is the mean of the difference in the number of threats identified by each subject.
- *Null hypothesis* (H_{0Cov}) = There is no significant difference in identifying types of threats using the proposed misuse case catalogue-based guideline and without using the proposed guideline. It can be formulated as H_{0Cov}: $\theta_{diff} = 0$, where θ_{diff} is the mean of the difference in the number of types of threats identified by each subject.
- *Alternative hypothesis* (H_{1Cov}) = There is a significant difference between identifying types of threats using the proposed misuse case catalogue-based guideline and without using the proposed guideline. It can be formulated as H_{1Cov}: $\theta_{diff} > 0$, where θ_{diff} is the mean of the difference in the number of types of threats identified by each subject.

3.3.1.2 Experimental Planning

Based on the hypotheses above, only one independent variable involved, which is threat identification used to identify threats in requirements, with two treatments: the proposed misuse case catalogue-based guideline and without the guideline. Two dependent variables need to be measured; effectiveness and coverage. To quantify the effectiveness of identifying threats, the number of threats or attacks identified by the participants at a given time was used. Furthermore, to quantify coverage, the number of types of threats found by the participants at a given time was used.

This experiment's subjects were Year 2 and Year 3 Computer Science students in the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak. Those students have already been equipped with knowledge on Unified Modelling Language (UML) diagrams, especially using the use case diagrams. However, none of these students got formal knowledge on security or quality as they have not attended any security or quality course in the university.

There were a few materials that were prepared for the experiment. The participants have been provided with an example of using the proposed catalogue-based guideline and without the proposed guideline, two tasks to be solved, Verb Pattern and Security catalogues, papers to list out the identified threats and types of threats, and a feedback form to express their opinion on what they think about the guideline.

Participants were given a use case diagram for a clinic appointment system and its descriptions in the first task. There are several use cases in the diagram related to the system, and the participants need to identify misuse case(s) for the diagram. For the second task, a use case diagram for an online shop system with its use case was given. Participants who needed to use the proposed guideline to solve the task were given catalogues for references.

This experiment applied Paired Comparison Design (1 factor with two treatments). The factors are threat identification, where the number of threats and types of threats will be identified. The treatments are "Without using guideline" and "Using catalogue-based guideline". The execution of this experimental design will be explained in the following subsection. Figure 3

shows the experimental design in the visual graphic.

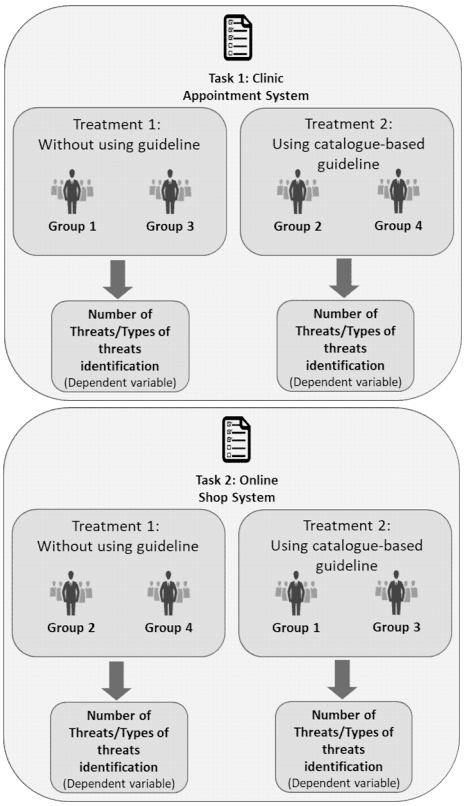

Figure 3: Experimental design

3.3.1.3 Experiment Execution

The first batch of the experiment was performed with 19 students from 10:00 am to 12:00 pm. We conducted an introduction lecture to introduce them to the misuse case technique and a tutorial on finding threats using a misuse case technique. Next, the participants were divided into Group 1 (9 students) and Group 2 (10 students). Both groups were given 30 minutes to find threats for each task. In Task 1, participants from Group 1 needed to find the threats without using the guideline, and for Group 2, they needed to find the threats using the proposed guideline. After they completed Task 1, they have to give an opinion on the technique they just used. Next, they took another 30 minutes to find threats in Task 2, where participants from Group 1 needed to find the threats by using the guideline, and participants from Group 2 needed to find threats without using the guideline. Opinion on the technique that they just used for Task 2 also has been recorded.

The second batch of the experiment was performed with 33 students. They were also divided into Group 3 (16 students) and Group 4 (17 students). They started from 10:00 am to 12:00 pm, following the same procedure as in the first batch.

3.3.1.4 Threats to validity

Validity evaluation is essential to make sure the results that we get from the experiment are valid. We identified two types of validity threats during experiments. The first type is threats to internal validity. These threats can affect the results without the knowledge of the researcher. The threats include:

- Learning effect - To avoid the learning effect from happening, Task 1 and Task 2 are developed based on a different domain. By doing this, the participants cannot refer to the knowledge gain from Task 1 when doing Task 2.
- Unwilling participants - Some participants were not interested in participating in this experiment and tried finishing it as early as possible. These types of participants will lead to inconsistency in the results and creating a few outliers. Here we used the box-plot technique to remove the outliers.
- Experimental fatigue - Task 1 and Task 2 were done continuously. Due to that, some of the participants may feel tired and unable to focus when completing Task 2. This problem can create biases towards the results where most probably Task 2 results might not be reliable. In order to avoid this problem, 5 minutes were given for them to rest before Task 2 started.

The second type is a threat to external validity. [15] defines the threat to external validity as an explanation of how we might be wrong in generalizing a particular study's findings. The threat to external validity for this experiment is that the experiment participants cannot represent the software developer as most of them are still studying and lack experience in use case diagram.

3.3.1.5 Results and analysis

This subsection presents the results of the experiment. Before doing the analysis, the unusable data need to be filtered out to avoid disturbing the calculation. Out of 52 participants, 3 of them were filtered out. The reason is that the answers provided were unrealistic, and it shows that the participants were not serious about completing the tasks. Only 49 data will be used in the calculation (refer to Table 4 and Table 5 for the refined data). The refined data were then tested to check for its normality using the box-plot technique.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: Refined data for threat identification.

		Threats Identification				
		without using guideline (Treatment 1)		Using catalogue-based guideline (Treatment 2)		
Group	Participants	Task 1	Task 2	Task 1	Task 2	Difference = Treatment 2-Treatment 1
G1	S1	0	-	-	7	7
G1	S2	0	-	-	4	4
G1	S3	0	-	-	5	5
G1	S4	3	-	-	4	1
G1	S5	2	-	-	6	4
G1	S6	1	-	-	8	7
G1	S7	0	-	-	6	6
G1	S8	1	-	-	6	5
G1	S9	2	-	-	9	7
G2	S25	-	1	1	-	0
G2	S26	-	4	5	-	1
G2	S27	-	4	5	-	1
G2	S28	-	3	1	-	-2
G2	S29	-	3	6	-	3
G2	S30	-	5	5	-	0
G2	S31	-	1	1	-	0
G2	S32	-	1	2	-	1
G2	S33	-	3	1	-	-2
G2	S34	-	3	1	-	-2
G3	S10	4	-	-	3	-1
G3	S11	3	-	-	5	2
G3	S12	0	-	-	1	1
G3	S13	0	-	-	1	1
G3	S14	2	-	-	3	1
G3	S15	1	-	-	1	0
G3	S16	1	-	-	1	0
G3	S17	3	-	-	3	0
G3	S18	0	-	-	2	2
G3	S19	0	-	-	2	2
G3	S20	1	-	-	2	1
G3	S21	0	-	-	7	7
G3	S22	2	-	-	2	0
G3	S23	3	-	-	4	1
G3	S24	4	-	-	2	-2
G4	S35	-	1	1	-	0
G4	S36	-	1	2	-	1
G4	S37	-	3	2	-	-1
G4	S38	-	4	4	-	0
G4	S39	-	1	2	-	1
G4	S40	-	3	2	-	-1
G4	S41	-	2	5	-	3
G4	S42	-	2	8	-	6
G4	S43	-	4	8	-	4
G4	S44	-	1	3	-	2
G4	S45	-	2	3	-	1
G4	S46	-	2	3	-	1
G4	S47	-	3	3	-	0
G4	S48	-	4	1	-	-3
G4	S49	-	4	1	-	-3

Table 5: Refined data for types of threats identification

		Types of threats identification				
		without using guideline (Treatment 1)		Using catalogue-based guideline (Treatment 2)		
Group	Participants	Task 1	Task 2	Task 1	Task 2	Difference = Treatment 2-Treatment 1
G1	S1	0	-	-	4	4
G1	S2	0	-	-	4	4
G1	S3	0	-	-	3	3
G1	S4	2	-	-	4	2
G1	S5	2	-	-	6	4
G1	S6	1	-	-	5	4
G1	S7	0	-	-	5	5
G1	S8	1	-	-	6	5
G1	S9	2	-	-	5	3
G2	S25	-	1	1	-	0
G2	S26	-	2	3	-	1
G2	S27	-	4	3	-	-1
G2	S28	-	2	1	-	-1
G2	S29	-	2	3	-	1
G2	S30	-	3	3	-	0
G2	S31	-	1	1	-	0
G2	S32	-	1	2	-	1
G2	S33	-	2	1	-	-1
G2	S34	-	3	1	-	-2
G3	S10	3	-	-	3	0
G3	S11	3	-	-	4	1
G3	S12	0	-	-	1	1
G3	S13	0	-	-	1	1
G3	S14	2	-	-	1	-1
G3	S15	1	-	-	1	0
G3	S16	1	-	-	1	0
G3	S17	3	-	-	3	0
G3	S18	0	-	-	2	2
G3	S19	0	-	-	2	2
G3	S20	1	-	-	2	1
G3	S21	0	-	-	5	5
G3	S22	2	-	-	2	0
G3	S23	3	-	-	3	0
G3	S24	4	-	-	2	-2
G4	S35	-	1	1	-	0
G4	S36	-	1	2	-	1
G4	S37	-	3	1	-	-2
G4	S38	-	3	2	-	-1
G4	S39	-	1	2	-	1
G4	S40	-	3	2	-	-1
G4	S41	-	2	2	-	0
G4	S42	-	2	4	-	2
G4	S43	-	3	3	-	0
G4	S44	-	1	3	-	2
G4	S45	-	2	3	-	1
G4	S46	-	1	3	-	2
G4	S47	-	3	3	-	0
G4	S48	-	3	1	-	-2
G4	S49	-	4	1	-	-3

A. GUIDELINE EFFECTIVENESS

In order to know whether the proposed guideline is better than the without guideline in terms of effectiveness in identifying threats, we do a hypothesis test.

The sample data shows that the corresponding sample means and the provided sample standard deviations are as in Table 6.

Table 6: Paired Samples Statistics for Effectiveness

		Mean	*n*	Std. Deviation	Std. Error Mean
Pair 1	Using Guideline	3.4694	49	2.30129	.32876
	Without Guideline	2.0000	49	1.44338	.20620

The results for paired sample test for effectiveness can be seen in Table 7. Then, we test our first hypothesis that we set earlier:

$H_{0Eff}: \mu = 0$ (1)

$H_{1Eff}: \mu > \mu_0$ (2)

This hypothesis corresponds to a right-tailed test, for which a t-test for two paired samples is used. The significance level, α, is the probability that the test statistic will fall in the critical region when the null hypothesis is true. For this test, we choose the common significance level, α = 0.05. Based on the information given, it is found that the critical value for this right-tailed test is t_c = 1.677. So, the rejection region for this right-tailed test is R = $t : t > 1.677$.

The test statistic, *t*, is computed by using the following formula:

$$t = \frac{\bar{D}}{S_D/\sqrt{n}} = \frac{1.46939}{2.67786/\sqrt{49}} = 3.841 \quad (3)$$

Since $t = 3.841 > t_c = 1.677$, we can conclude that the H_{0Eff} is rejected. That is means there is sufficient evidence shows that using the misuse case with proposed guidelines is more effective in identifying threats than using a misuse case without a guideline.

If we used the P-value approach, from the information given, $p = 0.00018$, and since $p = 0.00018 < 0.05$, we can get the same result, which is H_{0Eff} is rejected.Then, for the effect size, we used Cohen's *d* calculation:

$$d = \frac{\bar{D}}{S_D} = \frac{1.46939}{2.67786} = 0.5487 \quad (4)$$

As the effect size, *d*, is 0.5487, we can conclude that there is a medium effect in using the proposed guideline.

B. GUIDELINE COVERAGE

We used the same test to check whether the proposed guideline can identify more types of threats than without using the guideline.

From the sample data, using the same sample size, *n*=49, the corresponding sample means and the provided sample standard deviations are as in Table 8.

The results for paired sample test for coverage can be seen in Table 9. Then, we test our first hypothesis that we set earlier:

$H_{0Cov}: \mu = 0$ (5)

$H_{1Cov}: \mu > \mu_0$ (6)

Just like the first hypothesis, this hypothesis corresponds to a right-tailed test, for which a t-test for two paired samples is used. We choose the same significance level value, α = 0.05. Based on the information given, the critical value for this right-tailed test is t_c = 1.677. So, the rejection region for this right-tailed test is R = $t : t >$ 1.677.

The test statistic formula:

$$t = \frac{\bar{D}}{S_D/\sqrt{n}} = \frac{0.85714}{1.96850/\sqrt{49}} = 3.048 \quad (7)$$

Table 7: Paired Samples Test for Effectiveness

		Paired Differences					*t*	*df*	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Using Guideline – Without Using Guideline	1.46939	2.67786	.38255	.70022	2.23856	3.841	48	.000

Table 8: Paired Samples Statistics for Coverage

		Mean	*n*	Std. Deviation	Std. Error Mean
Pair 1	Guideline	2.5918	49	1.42768	.20395
	Non-Guideline	1.7347	49	1.20374	.17196

Table 9: Paired Samples Test for Coverage

		Paired Differences					*t*	*df*	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Guideline – Non-Guideline	.85714	1.96850	.28121	.29172	1.42256	3.048	48	.004

Since $t = 3.048 > t_c = 1.677$, we can conclude that the H_{0Cov} is rejected. That is mean there is sufficient evidence shows that by using misuse case with proposed guidelines can discover more types of threats than using misuse case without a guideline.

If we used the P-value approach, from the information given, $p = 0.00187$, and since $p = 0.00187 < 0.05$, we can get the same result which is H_{0Cov} is rejected.

Then, for the effect size, we used Cohen's *d* calculation:

$$d = \frac{\bar{D}}{S_D} = \frac{0.85714}{1.96850} = 0.4354 \quad (8)$$

As the effect size, d, is 0.4354, we can conclude that there is a small effect in using the proposed guideline.

3.3.1.6 Discussion

We used a one-tailed paired-samples t-test to compare the effectiveness of the misuse case in finding threats when using the catalogue-based guideline and without using the guideline. The mean numbers of threats identified were 3.4694 using the guideline and 2.0 without using a guideline. The hypothesis testing shows that there is sufficient evidence indicating that using the catalogue-based guideline is more effective than without using it, with the effect size at a medium level. The same situation also happens when comparing the ability to find types of threats. The mean for types of threats identified was 2.5918 when using the guideline, and 1.7347 without using the guideline, with a much smaller effect size. The effect sizes for both identifying threats and types of threats, which are medium and small, are considered acceptable due to several considerations. The first is due to the time slot that was used to complete the task. The participants were given a time slot to finish the task. Although the participants were not forced to complete the task, they still have a student mentality where they tried to finish it at the given time. This might give them the pressure to finish it quickly and thus affecting the effect size. Second, the participants are still new to the misuse case technique and the catalogue-based guideline, and the experience level dealing with this technique and guideline are not high. Even though the participants were given tutorial and discussion sessions before the experiment, it is still not enough to fully understand it. They might need to have extra time to be familiar with the technique and the guideline.

3.3.2 Expert review

After the empirical evaluation, expert reviews have been done. The purpose of conducting expert reviews is to verify the catalogue-based guideline's usability in eliciting security requirements. The experts also checked on the suitability of the techniques used in producing the catalogues.

3.3.2.1 Expert review hypothesis

The usability of the guideline was tested by testing the following hypothesis:

- *Null hypothesis* ($H_{0usefulness}$) = There is no significant difference in the usability of the guideline when producing misuse cases with a catalogue-based guideline and without the guideline. It can be formulated as H_0 $\mu_{Diff} = 0$.

- *Alternative hypothesis* ($H_{1usefulness}$) = There is a significant difference in the usability of the guideline when producing misuse cases with a catalogue-based guideline and without the guideline. This implies that the catalogue-based guideline is more usable than doing without the guideline. The hypothesis can be formulated as H_1 $\mu_{Diff} > 0$.

3.3.2.2 Design

This expert review applied Paired Comparison Design (1 factor with two treatments) where the experts as participants and the treatments are using a catalogue-based guideline (Treatment 1) and without using any guideline (Treatment 2). Two sets of problems were given, and the participants reviewed the usability of each treatment when applied to the problem. Usability was measured through two sets of USE questionnaires, one for each treatment (refer to Figure 4).

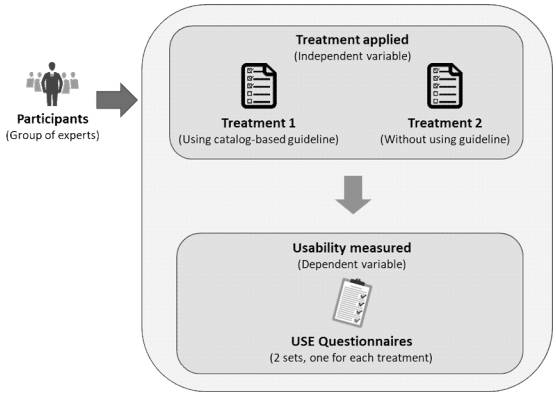

Figure 4: Design for expert review

3.3.2.3 Demographic of participants

Five experts were invited to do the review. All of them are adults aged 35 to 60 years old. Three of them are male. Most experts are experienced in software engineering, software quality, and software requirements with more than seven years of involvement in the fields. However, they have less experience in the software security field. Only one expert has more than seven years of experience in software security. In addition, two of the experts are also experienced software programmers. Table shows the demographics of the five experts.

Table 10: Demographics of the experts.

AGE				
Range of age	20-24	25-29	0-34	≥ 35
Number of experts	-	-	-	5 (100%)

GENDER		HIGHEST EDUCATION LEVEL			
Gender	Number of experts	Diploma	Degree	Master	PhD
Male	3	-	-	-	3
Female	2	-	-	-	2

FIELD EXPERIENCE				
Year / Domain	Nil	≤ 3	4-6	≥ 7
Software Engineering	1	-	-	4
Software Quality	1	-	1	3
Software Requirements	1	1	-	3
Software Security	1	1	2	1

3.3.2.4 Materials

There were a few materials that have been prepared for review. The experts have been provided with two USE questionnaires, two tasks, catalogues as a part of the guideline, and a feedback form.

Experts were required to identify misuse cases for an ATM system without using any guidelines in the first task. One use case diagram for an ATM system containing one actor and two use cases with its flow of events was prepared. For the second task, the experts were required to identify misuse cases for an online shop system with the catalogue-based guideline. An online shop system's use case diagram containing one actor and three use cases with its flow of events was prepared. The tasks were set on a different topic to avoid bias when solving the tasks.

3.3.2.5 Execution

During the expert review study, the experts were initially briefed on the study's purpose, the techniques used in producing the catalogues, and how to produce the misuse cases with the catalogue-based guideline and without the guideline. The experts were then invited to try producing the misuse cases with and without the guideline on two tasks. After completion of each task, the experts need to fill in the USE questionnaire to evaluate their experience when solving the task.

3.3.2.6 Threats to validity

Validity evaluation is essential to make sure the results that we get from the experiment are valid. We only identified one type of validity threat during experiments which are threats to internal validity. The threats include:

- Learning effect - To avoid the learning effect from happening, Task 1 and Task 2 are developed based on a different domain. By doing this, the experts cannot refer to the knowledge gained from Task 1 when doing Task 2.
- Experimental fatigue - Task 1 and Task 2 were done continuously. Due to that, the experts may feel tired and unable to focus when completing Task 2. In order to avoid this problem, 5 minutes were given for them to rest before Task 2 started.

3.3.2.7 Result and analysis

This subsection presents the result of the expert review.

Table 11: The data from expert review

Expert	With Guideline	Without guideline	Difference
1	178	90	88
2	208	111	97
3	154	117	37
4	190	156	34
5	203	90	113

Table 11 shows the data from the experts' reviews. The field "Expert" refers to the expert involved in the review. "With Guideline" refers to the total point given by the expert for the catalogue-based guideline. "Without guideline" refers to the expert's total point when producing the misuse cases without using a guideline, and "Difference" refers to the difference of points between "With Guideline" and "Without guideline".

A Shapiro-Wilk test has been done to check whether the data (difference of points) in Table 11 is normally distributed. Shapiro-Wilk test was chosen instead of Kolmogorov-Smirnov because the total number of participants was less than 30 samples. To be normally distributed, the difference point should have a significant value larger than 0.05. The difference point for the "With Guideline" and "Without Guideline" were normally distributed, as assessed by Shapiro-Wilk's test (Sig. = .230). Since the data are normally distributed, it is suitable for a paired t-test. The hypothesis testing section describes the result of the paired t-test. Table shows the normality test result on the difference of points for "With Guideline" and "Without Guideline".

Table 12: Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
difference	.253	5	.200*	.861	5	.230

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

A. DESCRIPTIVE STATISTIC

The USE questionnaire for the expert review consists of 4 criteria; usefulness, ease of use, ease of learning, and satisfaction. Figure 5 and Figure 6 below show the results of the questionnaire done by the experts.

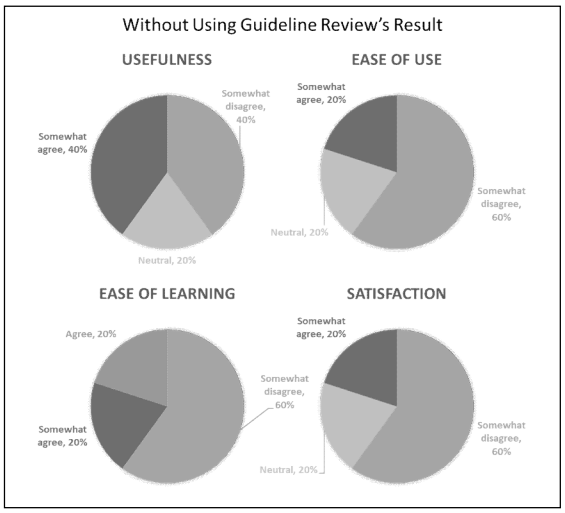

Figure 5: The results for "Without using guideline" review.

The pie chart illustrates the results from the USE questionnaire, which was defined by four main criteria; usefulness, ease of use, ease of learning, and satisfaction, obtained from the tasks reviewed by the experts. From the results, it can be seen that most participants felt somewhat disagree with the ease of use and satisfaction criteria when solving the task without using the guideline. Both criteria however change to agree and strongly agree when using the guideline while solving the task.

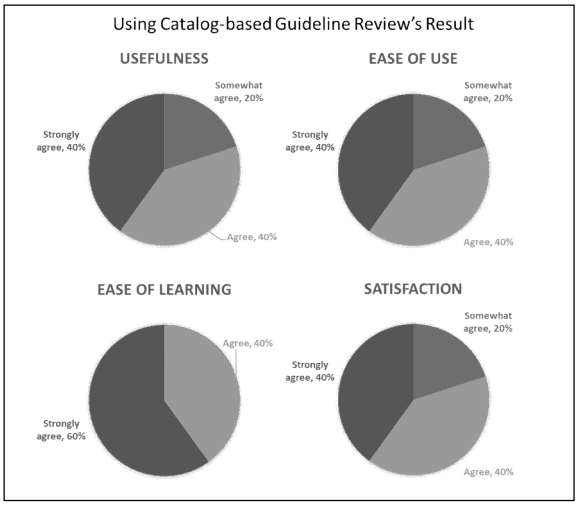

Figure 6: The results for "using catalogue-based Guideline" review.

The results obtained from the ease of learning criteria change from 60% somewhat disagree when solving the task without using the guideline to 60% strongly agree when the participants using the guideline while solving the task. The result for ease of use and satisfaction

criteria was both 60% somewhat disagree when solving the task without using the guideline, then positively change to 40% agree and 40% strongly agree when using the guideline. The proportion of expert reviews for ease of learning criteria also shows significantly positive results, from 60% somewhat disagree when not using the guideline to 60% strongly agree when using the guideline to solve the task. The overall result shows that the catalogue-based guideline is more useful, easy to use, easy to learn, and more satisfying than without a guideline when solving a task.

B. HYPOTHESIS TESTING

In order to know whether the proposed guideline is more usable than without a guideline, we do hypothesis testing using a paired t-test.

From the sample data, the corresponding sample means and the provided sample standard deviations are as in Table 13.

Table 1: Paired Samples Statistic for Usability

		Mean	*n*	Std. Deviation	Std. Error Mean
Pair 1	Misuse case with guideline	186.60	5	21.652	9.683
	Misuse case without guideline	112.80	5	27.050	12.097

The score differences and the degrees of freedom, *df*, can be referred to in Table 14. Standard deviation, S_D, is computed using this formula:

$$s_D = \sqrt{\frac{\sum(x_{diff}{}^2) - \frac{(\sum x_{diff})^2}{n}}{n-1}} = 36.107 \quad (9)$$

Then, we test our first hypothesis that we set earlier:

$$H_{0Usefulness}: \mu = 0 \quad (10)$$
$$H_{1Usefulness}: \mu > \mu_0 \quad (11)$$

This hypothesis corresponds to a right-tailed test, for which a t-test for two paired samples is used. The significance level, α, is the probability that the test statistic will fall in the critical region when the null hypothesis is true. For this test, we choose the common significance level, α = 0.05. Based on the information given, it is found that the critical value for this right-tailed test is t_c = 2.1318. So, the rejection region for this right-tailed test is R = t : t > 2.1318.

The test statistic, t, is computed by using the following formula:

$$t = \frac{\bar{D}}{S_D/\sqrt{n}} = \frac{73.8}{36.107/\sqrt{5}} = 4.570 \quad (12)$$

Since t = 4.570 > t_c = 2.1318, we can conclude that the $H_{0Usefulness}$ is rejected. That means there is sufficient evidence that shows using the misuse case with proposed guidelines is more usable in identifying threats than using a misuse case without a guideline.

If we used the P-value approach, from the information given, p = 0.00513, and since p = 0.00513 < 0.05, we can get the same result, which is $H_{0Usefulness}$ is rejected.

3.3.2.8 Discussion

This expert review study used descriptive statistics and hypothesis testing using a one-tailed paired-samples t-test to analyze the questionnaires' data.

Based on the participants' demographic, the experts' distribution shows that they were appropriate to evaluate the catalogue-based guideline from the software developer's perspective. Four of them have adequate knowledge of software engineering, software quality, and software requirements, and one of them is an expert in software security. From the USE questionnaire results, most experts agreed that using catalogue-based guideline is better than without using guidelines in terms of usefulness, ease of use, ease of learning, and satisfaction. According to the experts, the catalogue-based guideline helps them identify the threats and types of threats easier than without any guideline. The catalogue-based guideline is also quite simple, thus making it easy to learn. The experts agreed that it is much better to develop a catalogue-based guideline to become an assisting software tool for software developers.

From the hypothesis testing, we found that using misuse case with the proposed guideline is more usable (mean= 186.60, SD= 21.652) than without using the guideline (mean= 112.80, SD=27.050). If we look at the mean difference between these two treatments, using the proposed guideline got a mean increase of 73.8 points with

Table 24: Paired Samples Test for Usability

		Paired Differences					*t*	*df*	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Misuse case with guideline – Misuse case without guideline	73.800	36.107	16.147	28.968	118.632	4.570	4	.010

a 95% confidence interval [28.968, 118.632] more usable than without using the guideline. The results showed that the experts agreed that using the catalogues can help software developers identify more threats and types of threats/attacks. The experts also suggested that the Security Catalogue needs to be appropriately arranged, making it easier for a software developer to use the catalogue.

4. CONCLUSIONS

In this paper, we experimented with a misuse case with the catalogue-based guideline to identify threats and types of threats. The empirical evaluation shows sufficient evidence that the catalogue-based guideline is more effective in identifying threats and types of threats than without using guidelines. We also asked a few experts to review the usability of the catalogue-based guideline in assisting misuse cases to identify threats and types of threats. The results show that the proposed guideline is more usable in identifying threats than without the guideline.

Although the experts gave positive results on the usability of the guideline, the experiment needs to be replicated using the actual software developer to get a more reliable result. The experiment execution arrangement also needs to be appropriately planned, such as giving more time to understand the misuse case before the experiment to decrease the threats to validity.

In future work, we plan to implement the guideline into a tool that can identify the threats from the catalogue. We might need to implement boilerplates to produce better output.

ACKNOWLEDGEMENT

We would like to thank the Ministry of Higher Education Malaysia and Universiti Malaysia Sarawak (UNIMAS) for the financial support and a special thank to Universiti Putra Malaysia (UPM) for allowing us to complete this research.

REFERENCES:

[1] Sindre, G., Opdahl, A.L., "Eliciting Security Requirements with Misuse Cases", *Requirements Eng*, 10, 2005, pp. 34–44.

[2] Saeki M, Kaiya H., "Security Requirements Elicitation using Method Weaving and Common Criteria, *International Conference on Model Driven Engineering Languages and Systems*, Springer, Berlin, Heidelberg, Sep 28, 2008, pp. 185-196.

[3] Diallo, M. H., Romero-mariona, J., Sim, S. E., & Richardson, D. J., "A Comparative Evaluation of Three Approaches to Specifying Security Requirements", *12th Working Conference on Requirements Engineering: Foundation for Software Quality*, 2006, pp. 2–7.

[4] Gandotra, V., Singhal, A., & Bedi, P., "Identifying Security Requirements Hybrid Technique", *4th International Conference on Software Engineering Advances, ICSEA 2009, Includes SEDES 2009: Simposio Para Estudantes de Doutoramento Em Engenharia de Software*, 2009, pp. 407–412.

[5] Tøndel, I. A., Jensen, J., & Røstad, L., "Combining Misuse Cases with Attack Trees and Security Activity Models", *5th International Conference on Availability, Reliability, and Security (ARES 2010)*, 2010, pp. 438–445.

[6] Daramola, O., Sindre, G., & Stalhane, T., "Pattern-based Security Requirements Specification using Ontologies and Boilerplates", *2nd IEEE International Workshop on Requirements Patterns (RePa 2012)*, 2012, pp. 54–59.

[7] Yoo, Sang Guun, Hugo Pérez Vaca, and Juho Kim, "Enhanced Misuse Cases for Prioritization of Security Requirements", *In Proceedings of the 9th International Conference on Information Management and Engineering*, 2017, pp. 1-10.

[8] Ansari TJ, Pandey D., "An Integration of Threat Modeling with Attack Pattern and Misuse Case for Effective Security Requirement Elicitation",

International Journal of Advanced Research in Computer Science, 8(3), 2017.

[9] Khamaiseh S and Xu D., "Software Security Testing via Misuse Case Modeling", *In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, IEEE, 2017, pp. 534-541.

[10] El-Attar M and Nasser N., "Refactoring Misuse Case Diagrams using Model Transformation", *InENASE*, 2019, pp. 249-256.

[11] Coss, D., & Samonas, S., "The CIA Strikes Back: Redefining Confidentiality, Integrity and Availability in Security", *Journal of Information System Security*, 10(3), 2014, pp. 21–45.

[12] Fabian, B., Gürses, S., Heisel, M., Santen, T., & Schmidt, H., "A Comparison of Security Requirements Engineering Methods", *Requirements Engineering*, 15(1), 2010, pp. 7–40.

[13] Kamalrudin, M., Hosking, J., & Grundy, J., "Improving Requirements Quality Using Essential Use Case Interaction Patterns", 2011, 531.

[14] Robert J. Ellison, "Attack Trees", Software Engineering Institute, Carnegie Mellon University, 2005.

[15] Karpati P, Redda Y, Opdahl AL, Sindre G., "Comparing attack trees and misuse cases in an industrial setting", *Information and Software Technology*, 56(3), Mar 1, 2014, pp. 294-308.

[16] Trochim, W. M., & Donnelly, J. P., "The Research Methods Knowledge Base (Vol. 2). Cincinnati: OH: Atomic Dog Publishing, 2001.