Experimental Comparison of Two Safety Analysis Methods and its Replication

Jessica Jung[1], Kai Hoefig[2], Dominik Domis[3], Andreas Jedlitschka[1], Martin Hiller[4]

[1] Fraunhofer IESE
67663 Kaiserslautern, Germany
{name.surname}@iese.fraunhofer.de

[2] University of Kaiserslautern
67663 Kaiserslautern, Germany
hoefig@cs.uni-kl.de

[3] ABB Corporate Research, Industrial Software Systems
Ladenburg, Germany
dominik.domis@de.abb.com

[4] Cassidian
89077 Ulm, Germany
martin.hiller@cassidian.com

***Abstract*—(Background) Empirical Software Engineering (SE) strives to provide empirical evidence about the pros and cons of SE approaches. This kind of knowledge becomes relevant when there is the question of whether to change from a currently employed approach to a new one. An informed decision is required and becomes even more important in the development of safety critical system. For example, for safety analysis of safety-critical embedded systems, methods such as Failure Mode and Effect Analysis (FMEA) and Fault Tree Analysis (FTA) are used. With the advent of model-based development of systems and software, the question arises, whether safety engineering methods should also be adopted. New technologies such as Component Integrated Fault Trees (CFT) come into play. Industry demands to know the benefits of these new methods over the established ones, such as Fault Trees (FT). (Methods) For the purpose of comparing CFT and FT with regard to safety analysis methods' capabilities, such as quality of results and participants view on the methods' consistence, clarity, and maintainability, we designed a comparative study as a controlled experiment using a within-subject design. The experiment was run with seven academic staff members working towards their PhD. The study was replicated with eleven domain experts from industry. (Results) Although the analysis of the tasks' solutions showed that the use of CFT did not yield a significantly different number of correct or incorrect solutions, the participants rated the modeling capacities of CFT higher in terms of model consistency, clarity, and maintainability. (Conclusion) From this first evidence, we conclude that CFT have the potential to be beneficial for companies looking for a safety analysis approach for projects using model-based development.**

Keywords—model-driven development; safety analysis methods; fault trees analysis; component integrated fault trees; avionics; experiment; replication.

I. INTRODUCTION

Empirical Software Engineering (SE) strives to provide evidence about the advantages and disadvantages of SE approaches. This kind of knowledge becomes relevant when there is the question of whether to change from a currently employed approach to a new one. In avionic and other industries with software-intensive and safety-critical systems, model-based and model-driven development (MBD and MDD) approaches are more and more often used to handle the increasing cost and complexity of these systems. However, for new model-based approaches, experiences or empirical evidence are most often missing, which hampers their adoption in industry. This is particularly true for approaches that address the safety of avionic systems.

Avionics systems have to be certified by airworthiness authorities according to international standards such as ARP 4754 [1] and [2] before they can be installed and operated in an aircraft [3]. In order to get an avionics system certified, the company developing and manufacturing this system needs to provide the airworthiness authorities with evidence that the system is safe. One important step in creating this evidence is a safety analysis showing that the risk, i.e., the probability and severity, of hazards potentially caused by the system is below acceptable boundaries [4].

Such safety analyses consist of two main phases, (1) building the model representing the system from a safety perspective and (2) the analysis that is often performed with tool support. In our research, we focus on the modeling phase.

As of today, Fault Trees (FT) [5] are a safety analysis technique widely used in industry for modeling and analyzing the propagation of faults through the system, for identifying which hazards they cause, and for calculating the resulting hazard occurrence probability [6].

For safety analyses of model-based development approaches, a whole series of new analysis techniques and adaptations of existing ones have been proposed for substituting traditional fault trees. These model-based development approaches can be divided into three categories: Semantic Transformation, Fault Injection, and Failure Logic Modeling. Semantic Transformation annotates design models such as the Unified Modeling Language (UML) or the Systems Modeling Language (SysML) with additional safety information such as stereotypes, tags, or constraints. Afterwards it identifies patterns and transforms them into parts of the safety analysis model [7]. Fault Injection uses a formal design model such as Rich Components for analyzing the failure behavior of a system and generating the safety model automatically based on the results of a model checker [8].

Parts of this work were funded in the context of the project SPES2020 and SPES2020_XTCore, funded by the German Ministry of Education and Research under grant number 01IS0804 and 01IS12005E.

Failure Logic Modeling annotates the components or modules of a design model with modular failure models, which are composed automatically into the failure or safety model of the entire system based upon the system structure [9].

None of them is widely used in industry; not least due to the fact that they lack a thorough empirical evaluation The effectiveness and advantages of most of these approaches have only been argued or evaluated in single case studies..

To systematically provide first evidence, we designed a controlled experiment for the first phase of the safety analysis. Taking into account the request for studies with practitioners, we first performed it in academia (study 1) and replicated it with a sample from the avionics industry (study 2).

Together with our industrial partner, Cassidian, we decided to compare the currently used FT with Component Integrated Fault Trees (CFT) [10][11][12], a Failure Logic Modeling approach, in the analysis of SysML block diagrams. CFT are fault trees with a module concept for integrating them with the components of the design model. Due to their apparent similarity to the system design model, CFT promise to facilitate the understanding of the engineers, make it easier to maintain consistency between system design and fault tree, and to reduce the complexity of the analyses through a clear graphical approach.

In the experiment, participants had to work on four modeling tasks with the CFT as well as the FT methodology. Then the quality of the participants' task results and their personal opinion about the two tasks captured with a questionnaire after each task has been analyzed. The results of both studies indicate that CFT does not significantly produce a higher number of correct but also of incorrect results. Participants from both studies rated core qualities such as consistency, clarity and maintainability higher, but only in study 1 a significant result was achieved for clarity.

This paper is structured as follows. In Section 2, we present the background of our research, i.e., other studies that compared samples and the approaches applied in the experiment. In Section 3, we present the two safety analysis methods. In Section 4 we first discuss the research questions, derive hypotheses and then elaborate on the study design, measurement instruments, and data analysis. In Section 5, we present the results of study 1 followed by the results of the replication (study 2) in Section 6. In Section 7 we compare the replication to the original study, present threats to validity and report on lessons learned from the replication. Section 8 summarizes the work and gives an outlook to future work.

II. Related Work

Although empirical studies have been published on safety analysis approaches, controlled experiments are rarely reported. Stålhane and colleagues published a set of controlled experiments with students [13][14][15][16] comparing different approaches in the area of security and safety. They were specifically interested in the effect the approach had on the subjects' performance, i.e., the number of hazards the subjects were able to identify using the methods based on the given requirements. In our experiment, for CFT, the focus was on the failure propagation of internal and external failures through the system to the hazards, which may occur at the system boundaries (e.g., loss of an aircraft engine). Thus, the hazards must be known before a fault tree analysis can start, which is a different situation. In their experiments, Stålhane et al. also investigated perceived ease-of-use, perceived usefulness, and intended use by using the technology acceptance model (TAM) [17]. In an experiment with 42 students [13], Stålhane and Sindre compared misuse cases (MUC) to Failure Mode and Effects Analysis (FMEA) in analyzing use cases. The aim of the experiment was to identify failures of the user; however, as for CFT, the focus of FMEA is on failure propagation (causes and effects) through the system and not primarily on failures of the user. This difference is confirmed by the results: MUC is better than FMEA for analyzing failure modes related to user interactions; FMEA is better than MUC for analyzing failure modes related to the inner workings of the system; MUC creates less confusion and is generally easier to use than FMEA. In a comparable setting [14], the authors compared MUC based on use-case diagrams to those based on textual use cases (TUC). MUC were developed for security analysis; here they are used for hazard identification with respect to safety. The results of the experiment with 52 students show that text produces better results because of more detailed information. Two further experiments are reported comparing system sequence diagrams (SSD) and TUC as specification techniques to determine their respective strengths and weaknesses with regard to finding and documenting hazards [15]. The experiment was split between two locations; 29 students applied TUC at location A and 10 students applied SSD at location B. The results show that TUC are better for identifying hazards related to the operation of the system, while SSD are better for the identification of hazards related to the system itself. In their most recent paper [16], the authors present an experimental comparison between TUC and SSD conducted with 48 students. The results partially confirm that TUC is better for identifying hazards with respect to human involvement and SSD is better with respect to inter-system aspects.

Opdahl and Sindre [18] describe two controlled experiments with students comparing attack trees and MUC, which were both developed for security analysis. However, attack trees are considered to be (very) similar to fault trees. The authors were interested in the effectiveness of the techniques, i.e., in the number of threats found, and in their coverage, i.e., in the types of threats found. Like Stålhane et al., they used TAM to measure the subjects' perception. They changed the setting from the first to the second experiment by giving a pre-drawn use-case diagram to the subjects of the latter. The main finding was that attack trees were more effective for finding threats, in particular, if no pre-drawn use-case diagram was available.

For the certification of safety-critical systems, traceability between safety requirements and architectural models (SysML) is important, in addition to safety analysis. In this context, Briand et al. [19] investigated the impact of the use of SysML design slices on inspectors' decision correctness and effort in a controlled experiment with 20 graduate students. The results

show a significant decrease in effort and an increase in the correctness and level of certainty of the decisions.

Pai and Dugan [20] empirically evaluated their Bayesian network (BN) model, relating object-oriented software metrics to software fault content and fault proneness, by using a public domain data set from a software subsystem.

Concluding, only a few studies are available that investigate techniques related to model-based safety analyses [13][14][15][16][18][19]. Most of them apply techniques from the area of security [13][14][15][16][18] to support threat identification. Only one of the studies evaluates a safety analysis method (FMEA) but applies it for a different purpose [13]. We did not find any experiment evaluating FT or CFT. Learning from those experiments, the assessment of participants' perception is an appropriate complement to objective metrics such as number of correct solutions or number of identified threats.

III. Technical part: Safety Analysis Methods

In the development of embedded systems such as avionic systems, model-based approaches such as UML and SysML are more and more often used. Fig. 1 shows an obfuscated and simplified excerpt of the SysML Block Diagram [7] used in the experiment. The *Block* consists of an input port *I1*, an output port *O1*, and four subcomponents (properties): The sensor *S* reads the input *I1* and sends its result via the two redundant channels *C1* and *C2* to the voter *V*. *V* selects one of the available signals and forwards this to the output port *O1*.

If one of the components causes an error or the signal was already erroneous at *I1,* the output *O1* becomes erroneous. This is modeled by the fault tree (FT) [5] in Fig. 2a) with the top event *O1 Err*. The tree consists only of Boolean Or-gates and basic events (BE) modeling the erroneous state of a component or of the in-port *I1* such as *I1 Err*. Fig. 2b) shows the fault tree of the top event *O1 Loss*, i.e., an omission of the signal at the output port *O1*. *O1 Loss* is caused by a loss of *V* (*V Loss*), by a loss of both channels (*C1 Loss* and *C2 Loss*), by a loss of *S*, or by loss of *I1*. Because the signal of both channels must be lost to cause *O1 Loss*, an AND-gate is used to connect it's both sub-trees. In fault trees, BE such as *S1 Loss* are modeled twice, because two paths exists from the BE to the top event.

In contrast to FT, Component Integrated Fault Trees (CFT) [10][11][12] are not independent of, but formally and visually integrated with the SysML Block Diagram from Fig. 1. Fig. 3 shows in the background the SysML Block Diagram of a *Channel* as well as its ports *i1* an *o1* (in Fig. 1, *C1* and *C2* are instances of *Channel*). In the foreground, Fig. 3 shows the CFT of *Channel,* which consists beside the same Boolean gates and BE as the FT of input and output failure modes. Output failure modes are modeled as dark-filled triangles at the top and input failure modes as non-filled at the bottom of a CFT. The input failure modes *i1_err* models an erroneous input signal at *i1* and *i1_loss* models the omission. Both input failure modes are visually and formally integrated with *i1* by the dashed lines between them. The same is true for the output failure modes *o1_err* as well as *o1_loss* and *o1*. In addition, the CFT is formally liked to *Channel* by the dashed line on the right of Fig. 3. In this way, the CFT, its gates, and it's BE are unambiguously assigned to *Channel* and its in-/output failure modes are formally linked to the ports of *Channel.*

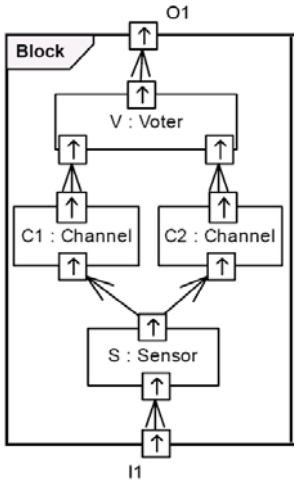

Fig. 1. Example SysML Block Diagram

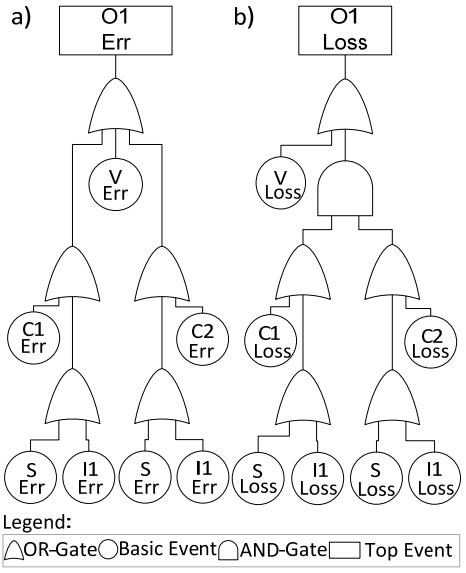

Fig. 2. a) FT with top event "*O1 Err*" for the example Block and b) FT with top event "*O1 Loss*".

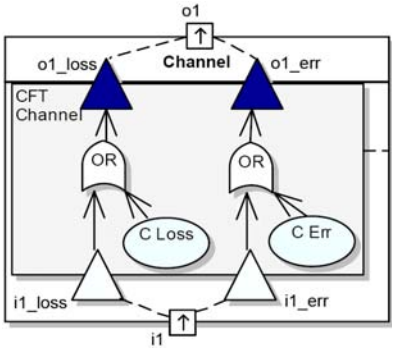

Fig. 3. SysML Block Diagram and CFT of a Channel

Fig. 4 shows the overview CFT of *Block*, which consists of the CFTs of its components *V, C1, C2*, and *S* as well as the input failure modes *I1 Err* and *I1 Loss* of *I1* and the output failure modes *O1 Err* and *O1 Loss* of *O1*. The CFT of *Block* describes the same Boolean trees as the FT in Fig. 2, but preserves also the modular component or block structure of the SysML model in Fig. 1. This modular structure and the visual links between the CFT and the SysML blocks, facilitate the traceability between functional and safety model and increase the clarity of the CFT compared to the separated and formally independent FT. Clarity means that the graphical presentation of the system supports keeping an overview and facilitates understanding. In addition, the better traceability increases the consistency between safety and SysML model and simplifies the maintainability of the safety model. Consistency means that the system design model and the fault tree describe the same system and version, for example, and do not contradict each other. Maintainability means that changes in the system (model) can be easily updated in the FT/CFT. For example, what changes in the safety model, if instead of two redundant channels only one or three are used in the SysML model? In the CFT of Fig. 4, the number of *Channel* CFT as well as the CFT of *V* has to be adapted accordingly. In the FT of Fig. 2, the entire tree has to be traversed for identifying the corresponding gates and BE. This becomes a very complex task for larger trees. In order to answer the question whether CFT actually perform better with respect to clarity, consistency and maintainability than FT, we set up the empirical experiment described in the next section.

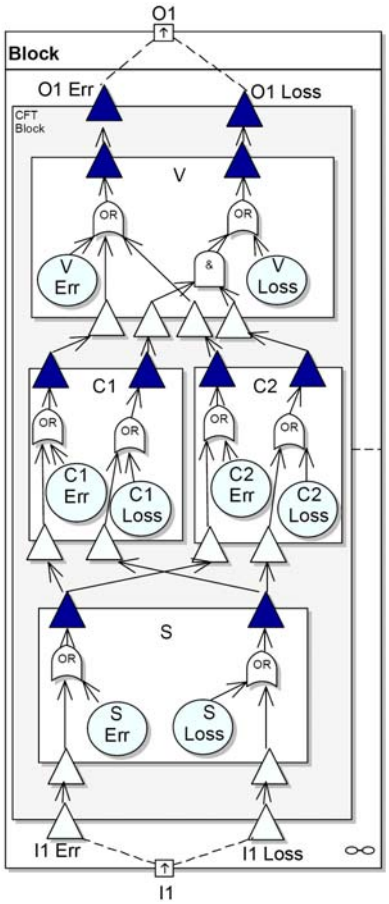

Fig. 4. Overview CFT of Block

IV. Experimental Methodology

Our work aims at exploring how a model-based safety analysis method (CFT) behaves in comparison to the established methodology (FT). For this purpose, we conducted an experiment and its replication based on the following goals and hypotheses.

A. Study Goal and Hypotheses

CFT and the system development model use the same graphical model, therefore we assume that safety and system engineers are supported in building a mental representation [21][22][23] of the relationship between the system model and the corresponding CFT. In addition, only a reduced part of the entire safety analysis model has to be investigated to identify inconsistencies between system model and fault tree when using CFT. Thus, we want to research if CFT are less error-prone and if participants will favor CFT over FT. Therefore, we specified two research questions (RQ) upon which the empirical study was built:

1) RQ1: Will the application of the CFT yield the same quality of the resulting safety model compared to a model built with FT?

This question relates to the performance of the techniques and will be investigated by analyzing the quality of the participants' results in modeling tasks performed with the two techniques. We define quality as correctness with three instances: number of *correct solutions* (all task solutions that are completely correct or follow the correct logic), number of *incorrect solutions* (all attempts of task solutions that do not follow the correct logic, use the wrong system element, or are incomplete), and number of *missing solutions* (all tasks that were not worked on by the participants, meaning that no marks or attempts to solve the problem are visible). For this comparison, we state the following hypothesis:

- H_1: When using CFT, participants will produce results with a different level of quality compared to when they use FT.

H_{11}: # correct solutions CFT $\neq$ # correct solutions FT

H_{12}: # incorrect solutions CFT $\neq$ # incorrect solutions FT

H_{13}: # missing solutions CFT $\neq$ # missing solutions FT

The corresponding null-hypotheses are:

$H0_{11}$: # correct solutions CFT $=$ # correct solutions FT

$H0_{12}$: # incorrect solutions CFT $=$ # incorrect solutions FT

$H0_{13}$: # missing solutions CFT $=$ # missing solutions FT

2) RQ2: Is CFT perceived different than FT with regard to consistency, clarity, and maintainability?

To answer this question, the participants were asked to provide their opinion in a questionnaire after each modeling task. Subjective perception focuses on three important fault tree

qualities: *consistency* (the system design model and the fault tree describe the same system), *clarity* (the graphical presentation of the system supports keeping an overview of the system), and *maintainability* (changes in the fault trees due to updated system models are easy to administer). Ratings are given on a scale from one (lowest) to five (highest). We state three hypotheses:

- H_{21}: When using CFT, consistency between the system description and the safety analysis model is perceived different when using FT.

  H_{21}: μCFT ≠ μFT

  $H0_{21}$: μCFT = μFT

- H_{22}: When using CFT, the clarity of safety analysis models developed with CFT is perceived different when using FT.

  H_{22}: μCFT ≠ μFT

  $H0_{22}$: μCFT = μFT

- H_{23}: When using CFT, the maintainability of safety analysis models developed with CFT is perceived different when using FT.

  H_{23}: μCFT ≠ μFT

  $H0_{23}$: μCFT = μFT

B. Study Design

The study was designed as an experiment under laboratory conditions not to exceed a given time frame of 4 hours. To investigate the hypotheses, various preliminary considerations were implemented in the experimental set-up.

1) Study Preparation

a) To secure a basic level of knowledge of the two methods, a tutorial session about SysML, CFT, and FT is provided before the actual study starts. The tutorial describes the two methods and explains the rationale behind the methods as well as the relationship between CFT, system model, and FT.

b) To provide a deep cognitive processing of the two methods, both methods should be applied by the participants. In total, four tasks were defined by three employees of Cassidian (who are working in the safety analysis department of avionics) according to typical activities in their everyday work. The four tasks are presented in two versions: version "a" for CFT and version "b" for FT (for reasons of comparability, the tasks for applying either the CFT or the FT method are identical). The participants had to work on both task versions in random order, so that a total of eight tasks were processed in the study (resulting in a within-subject design [24]). To avoid ordering effects such as practice, learning effects, or fatigue, the task sequence, i.e., the application of either the CFT or the FT method, was randomized for each participant, resulting in a different task order for each individual [24]. For randomization we generated randomization sets, which were created by randomly combining task number (1-4) and method (a = CFT, b = FT; e.g., Set 1: 1b 1a 4a 2b 3b 3a 4b 2a). In task 1, the participants are asked to include two missing elements ("basic events") in the given model and perform the corresponding changes in the model. In task 2, a new additional functionality of the system is to be included in the given error analysis model. For the solution of task 3, the participants have to implement an additional component in the system and transfer all necessary changes in the corresponding fault trees, respective component fault trees. In task 4, an additional analog input component is to be added to an already existing component in the model.

c) The application of safety analysis methods is normally done with PC supported modeling tools. As effects caused by the usability of these tools are likely to influence the subjective rating of the safety analysis method, the study is implemented for paper & pencil use. Standardized instructions describe the usage and processing of the provided materials. For working on the tasks, the participants are instructed to draw in missing elements directly in the diagrams provided and cross out non-essential elements. Blank pages are provided to redraw elements. In addition, the system model is provided during the processing of the tasks as a reference document. The challenge in the processing of each task is first to identify the task-relevant diagrams out of all diagrams provided for the solution and second, to select and implement the correct changes (e.g., adding items, deleting items) in all necessary spots. During the whole study, no group work is allowed.

d) For future reuse of the study materials and the expected internationality of participants in the information science domain, all materials (instructions, tutorial, task descriptions, system model and questionnaires) are provided in English.

e) To minimize experimenter effects and to promote replicability of the study, all instructions and materials are standardized. Therefore the study process and the accompanying materials are developed in a way that no interaction between participants and experimenter is necessary after the initial introduction to the experiment. All necessary documents (except the pre-questionnaire) are combined into one document, the "*test booklet*". This document is self-describing, e.g., explains how the questionnaires should be completed and that the tasks are to be processed in the order provided, and it contains the already mentioned tasks to be solved by applying either the CFT- or FT-methodology. In the test booklet each task is coupled with a short questionnaire assessing the subjective assessment of the dependent variables consistency, clarity and maintainability. A total of eight tasks have to be processed, four for each methodology.

2) Study Procedure

The standardized study process is described as follows. 1) The study participants are greeted by the experimenter with the help of a standardized instruction. This includes informing the subjects about the purpose and procedure of the study, a description of the documents to be used in the study as well as obtaining informed consent of the participants. 2) A first questionnaire (*pre-questionnaire*) covering occupational background and previous experience with safety analysis methods is handed out for processing. 3) This is followed by the *tutorial*. 4) Thereafter the *test booklets* are handed to the participants. They finish with a *post-questionnaire* consisting of quantitative and qualitative questions for the subjective assessment of CFT. 5) To wrap up the study, an *open feedback*

TABLE I. QUESTIONNAIRE ITEMS TO ASSESS PERCEIVED CONSISTENCY, CLARITY AND MAINTAINABILITY

Scale Perceived Consistency	Scale Perceived Clarity	Scale Perceived Maintainability
1. I am sure that I was able to transfer the modifications from the system model completely to the <CFT/FT>.	5. Because of the graphical representation of <CFT/FT> it was easy for me to keep the overview of the failure logic.	9. It was easy for me to implement the modifications.
2. It was easy for me to transfer modifications in the system model to the <CFT/FT>.	6. The relationship between the <CFT/FT> and the system is easy for me to comprehend.	10. I was able to make the modifications with minor effort.
3. I was able to identify the locations in the <CFT/FT> that needed to be involved for doing the modifications.	7. The <CFT/FT> methodology helped me to keep the overview of the failure logic.	11. I was able to reuse a lot from the existing model during the modifications.
4. I am sure that I was able to identify all the involved locations.	8. The <CFT/FT> supported me during the accomplishment of the tasks.	-
5-point ordinal Likert Scale: 1=strongly disagree, 2=disagree, 3=neither agree nor disagree, 4=agree, 5=strongly agree		

session and the opportunity for exchange and discussion among the participants is provided. Information gathered in the feedback session as well as the post-questionnaire is not analyzed in this paper.

The study design was piloted three times, twice at Cassidian (n=2) and once at Fraunhofer IESE (n=4), to check whether the materials were understandable and manageable for the respondents and whether the process and the actual timeframe (3.5h) coincided with prior planning. In study 1 as well as study 2, participation was voluntary and participants received no compensation for participating.

C. Measurement Instruments

To answer the two research questions, we collect the data resulting from the task processing and record the participants' opinions about the two methods with the help of questionnaires. In the following, these measurement instruments are described.

1) Task results analysis: collecting objective data

To analyze the objective data provided by the task results, a classification scheme was developed. Using this classification scheme, all task solutions can be uniquely assigned to a specific category. The categories are as follows: (1) Correct solution: All task solutions that are completely correct or follow the correct logic are assigned to this category. (2) Incorrect solution: All attempts of task solutions that do not follow the correct logic, use the wrong system element, or are incomplete are allocated to this category. (3) Task solution missing: This category includes all tasks that were not worked on by the participants, meaning that the participant left all task-related pages blank (no marks or attempts to solve the problem are visible). As a result of this coding scheme, the total number of correct, incorrect, and missing task solutions needed to compare CFT with FT is specified in a list.

2) Questionnaire: collecting subjective data

To assess participants' opinions about the two methods, a standardized questionnaire was developed. This questionnaire is to be answered immediately after completion of a task and before a participant moves on to the following task. So in total eight questionnaires had to be filled by each participant. The content of the questionnaire is based on the pre-established hypotheses: each property of the methods (clarity, consistency, maintainability) was transferred into several items, 11 items altogether. These 11 items (=statements) should be rated on an ordinal 5-point Likert scale (1=*strongly disagree* to 5=*strongly agree*) for each task [25]. Depending on the task type ("a" or "b"), the item contains either the term "CFT" or "FT". The original items are listed in TABLE I.

D. Data Analysis

Data analysis for RQ1 consists of two steps: (1) categorizing the results of the modeling task for both safety analysis methods and (2) comparing the two methods with respect to number of correct task solutions, incorrect task solutions, and missing task solutions. Data analysis for RQ2 involves (1) reliability testing of the three questionnaire scales, (2) merging the items of one scale into one variable (sum of the values divided by the number of items), (3) comparing the two methods with respect to the three computed variables. To test the hypotheses, we use the McNemar Test [26] for H_{11}-H_{13} and the Wilcoxon signed-rank test (non-parametric test for dependent samples) for H_{21}-H_{23} on a significance level of 0.95 (α = 0.05, two-tailed) [27]. In the following, we present the results and statistical analyses for the two research questions, first for study 1 (Section V), then for the replication (Section VI).

V. STUDY 1

A. Sample Description

In study 1, seven academic staff members of the University of Kaiserslautern (TU KL) participated (n=7; 3 female). All of them are computer scientists on their way to their doctoral graduation (PhD) with experience in safety from projects with industry. Five of them are working in the area of safety analysis, two in the field of software engineering. Besides these background variables, the participants were asked to self-assess their prior method knowledge regarding safety analysis, fault tree analysis, and system design in general, and regarding SysML in particular. They rated their degree of experience on an ordinal 5-point scale (1=*little experience* to 5=*a lot experience*) and in years working with these topics in the pre-questionnaire. TABLE II. shows the descriptive results. Overall, participants report lowest experience levels with SysML and a medium level of experience for the other three categories.

B. Results: Hypotheses testing

To answer RQ1, a total of 56 task solutions (7 participants * 4 tasks * 2 methods) were analyzed by applying the coding scheme described in Section IV. The total numbers of the three categories for the CFT and the FT methodologies are shown in TABLE III. FT has a higher number of correct solutions but also a slightly higher number of incorrect solutions compared to CFT. The number of missing task solutions when applying CFT is noticeable. However, when testing the differences with the McNemar Test (two-tailed) none of the comparisons revealed a significant result. When using CFT, the participants did not significantly produce more or fewer correct solutions (H_{11}). The same results emerge for the number of incorrect solutions (H_{12}) and the number of missing task solutions (H_{13}); again we retain the null hypothesis: when using CFT, the participants had an equal number of correct ($p = 0.375$), incorrect solutions ($p = 1.00$), as well as missing solutions ($p = 0.125$) compared to FT.

To investigate RQ2, the subjective opinion about the two methods with respect to the three qualities consistency, clarity, and maintainability is of interest. Each quality is represented with one scale consisting of several items (see TABLE I.). For further statistical analyses, a reliability check [28] of the three subscales is needed to merge the individual items of one scale in a scale-averaged value. Cronbach's α is one reliability measure and it measures the internal consistency of the items of one scale. Its value varies between 0 and 1 (values close to 1 indicate high scale reliability) [28]. Across all four tasks, Cronbach's α for the scale consistency measuring CFT is 0.80 and for FT it is 0.96. The scale clarity assessing CFT has a Cronbach's α of 0.85 and for FT 0.89. The scale maintainability has a Cronbach's α of 0.85 for CFT and 0.92 for FT. These results indicate a good to excellent reliability of the developed scales. Based on this result, the individual item values of one scale (e.g., item 1 to 4) were merged into a new variable (sum of the values divided by the number of items) representing the scale value for each participant. Then the Mean and Median values of the three new variables were assessed combined for the four tasks for each safety analysis method (CFT vs. FT). We further used these new variables for investigating H_{21}-H_{23}. The results for the Wilcoxon signed-rank test to examine H_{21}-H_{23} are displayed in TABLE IV. The participants experienced clarity significantly higher for CFT (Md=4.00) compared to FT (Md = 3.00), $z = -2.41$, $p < .05$, $r = -0.64$ (H03 can be rejected). Furthermore consistency as well as maintainability was experienced higher for CFT compared to FT, here a significant result was missed ($p > 0.05$). However, for all three comparisons of the safety analysis methods, we attain (close to) medium effect sizes[1] for consistency and maintainability and a large effect size for clarity (see TABLE IV.

[1] The effect size r of the Wilcoxon signed rank test is calculated by dividing the test statistic z by the square root of the number of observations. The absolute value of 0.1 is defined as small effect size, |0.3| as the threshold for medium and |0.5| for large effect sizes [29].

TABLE II. DESCRIPTIVES OF SELF-ASSESSED EXPERIENCE STUDY 1

Method	Descriptive Results			
	safety analysis	***fault tree analysis***	***system design***	***SysML***
Self-assessment on 5 point scale	Md = 3	Md = 3	Md = 3	Md = 1
Years working with these topics	M = 1.64 SD = 1.03	M = 1.57 SD = 1.21	M = 2.93 SD = 2.86	M = 0.43 SD = 0.78

Annotations: Md: Median, M: mean, SD: Standard Deviation.

TABLE III. NUMBER OF CORRECT, INCORRECT OR MISSING SOLUTIONS FOR STUDY 1 (RESEARCH QUESTION 1)

Method used	Results for Task 1-4			
	H_{11}: # correct solutions	***H_{12}: # incorrect solutions***	***H_{13}: # missing task solutions***	***total***
CFT	22	2	4	28
FT	25	3	0	28
Total	47	5	4	56
p[a]	0.375	1.00	0.125	-

[a] McNemar Test ($\alpha = 0.05$; two-tailed).

TABLE IV. RESULTS OF RESEARCH QUESTION 2 OF STUDY 1

	Results for Task 1-4 combined					
	H_{21}: Consistency		***H_{22}: Clarity***		***H_{23}: Maintainability***	
Method	CFT	FT	CFT	FT	CFT	FT
Mean (SD)	4.12 (0.43)	3.57 (0.93)	4.18 (0.37)	3.18 (0.51)	3.79 (0.81)	3.57 (0.94)
Md	4.00	4.00	4.00	3.00	4.00	3.00
z	-1.732		-2.410		-1.069	
p[a]	0.083		**0.016**		0.285	
r	-0.34		-0.64		-0.29	

Annotations: SD: Standard Deviation, Md: Median, z: test statistic, r: effect size.
[a] McNemar Test ($\alpha = 0.05$; two-tailed).

VI. REPLICATION – STUDY 2

In the replication, the same instructions, instruments and materials were used; no changes to the design, used artifacts, procedures, data collection methods and analysis techniques were applied. Also the same experimenters from study 1 lead the replication study. Different to study 1 is the sample and the location (the study was conducted at the company's site instead of the Fraunhofer IESE). Instead of an academic sample, a practitioners sample from the avionics domain participated.

A. Sample

In the replication (study 2), eleven employees of Cassidian participated (n=11; 1 female). Their educational background can be described as follows: eight are aerospace engineers, two are mathematicians, and one is physicist. They are mainly employed in the area of safety analysis (9 of 11 employees); two are working in the field of software engineering. Coinciding to study 1, participants were asked to self-assess their experience (see 0). Again, experience with SysML is rated lowest. Compared to the sample of study 1, the self-assessment for the other three categories is alike whereas the number of years working with these three topics is estimated higher.

B. Results

The results for RQ1 are similar to study 1. TABLE VI. shows, that again FT has a higher number of correct and incorrect solutions while CFT has a fairly larger number of

TABLE V. DESCRIPTIVES OF SELF-ASSESSED EXPERIENCE STUDY 2

Method	Descriptive Results			
	safety analysis	***fault tree analysis***	***system design***	***SysML***
Self-assessment on 5 point scale	Md = 3	Md = 3	Md = 3	Md = 1
Years working with these topics	M = 3.68 SD = 3.08	M = 3.36 SD = 3.20	M = 5.05 SD = 4.59	M = 0.23 SD = 0.41

Annotations: Md=Median, M=mean, SD=Standard Deviation.

TABLE VI. NUMBER OF CORRECT, INCORRECT OR MISSING SOLUTIONS FOR STUDY 1 (RESEARCH QUESTION 1)

Method used	Results for Task 1-4			
	H_{11}: # correct solutions	***H_{12}: # incorrect solutions***	***H_{13}: # missing task solutions***	***total***
CFT	30	4	10	44
FT	34	5	5	44
Total	64	9	15	88
p[a]	0.125	1.00	0.125	-

[a] McNemar Test ($\alpha = 0.05$; two-tailed).

TABLE VII. RESULTS OF RESEARCH QUESTION 2 OF STUDY 2

	Results for Task 1-4 combined					
	H_{21}: Consistency		***H_{22}: Clarity***		***H_{23}: Maintainability***	
Method	CFT	FT	CFT	FT	CFT	FT
Mean (SD)	3.64 (1.18)	3.39 (1.03)	3.48 (1.20)	3.11 (1.20)	3.68 (1.12)	3.50 (1.09)
Md	4.00	3.25	4.00	3.25	4.00	4.00
z	-1.289		-0.841		-0.680	
p[a]	0.197		0.400		0.496	
r	-0.27		-0.18		-0.14	

Annotations: SD: Standard Deviation, Md: Median, z: test statistic, r: effect size.
[a] McNemar Test ($\alpha = 0.05$; two-tailed).

missing solutions. The McNemar test revealed that none of the comparisons is significant. We retain the null hypothesis: CFT and FT do not differ in terms of correct ($p = 0.125$), incorrect ($p = 1.00$) and missing task solutions ($p = 0.125$).

To answer RQ2 we again first analyze the reliability of the used scales. Cronbach's α for the scale consistency is 0.94 for CFT and 0.97 for FT. The scale clarity takes values of 0.97 (CFT) and 0.98 (FT). For maintainability Cronbach's α is 0.92 in case of assessing CFT and 0.93 in case of FT. Again we computed one scale value out of the scale's items for each participant. Then we analyzed the new variable combined for task 1-4 and compared CFT against FT with the two-tailed Wilcoxon signed-rank test. The results (see TABLE VII.) show, that all of the hypotheses have to be rejected ($p > 0.05$); also the effect sizes are rather small ($r = -0.27$; $r = -0.18$; $r = -0.14$) compared to study 1. CFT is rated only slightly higher than FTs regarding consistency, clarity, and maintainability.

VII. DISCUSSION

A. Comparison of Study 1 and its Replication

The experiment was first performed with academic staff of TU Kaiserslautern with project experience in the safety domain (n=7) and then replicated with a practitioners sample from Cassidian (n=11). From study 1 to study 2 the only changes applied to the study was the used sample (and with probably minor effects the study location). Interestingly, the self-assessment of the sample of study 1 and the sample of study 2 does not discriminate between the academic and the practitioner's sample. In both samples the same self-assessment patterns appear, however, with a higher variance in the sample of study 2 for the areas safety analysis, fault tree analysis, and system design in general. Only the self-assessment in years indicates that the practitioners have a larger experience basis compared to the academics. For both samples SysML is a new methodology.

The results regarding the quality of the task results (RQ1) in study 2 are consistent to the results of study 1. The results of both studies showed that when applying CFT, the participants did not produce significantly more or fewer correct and incorrect solutions compared to FT. CFT do neither produce a larger number of correct solutions nor does the application of CFT result in a larger number of incorrect solutions. Also a remarkable, but not significantly larger number of missing solutions appear for CFT, which can be attributed to uncertainty in applying the CFT method.

Regarding the subjective assessment of the two methods, in study 1 CFT was rated higher than FT for consistency, clarity and maintainability, but only the comparison for clarity reached a significant level. The results of the replication slightly differ from the results obtained in study 1: the differences between CFT and FT have the same direction but are very small and far from a significant level. Concerning RQ2, the obtained effect sizes differ in their value between the two studies; the effect sizes obtained in study 1 reach values higher |0.29| (medium to high effect sizes) whereas in study 2 the effect sizes vary between |0.14| and |0.27|, indicating rather small effect sizes. The difference in the results might derive from the used sample. Several explanations are possible: it might be possible, that practitioners might take a rather critical or opposing position against new methods, because it is not (yet) clear, if these new methods are beneficial and they want to avoid high efforts for learning these methods. Therefore the technology effect would be systematically underestimated. In contrast, academics are rather open for new experiences and new methods. After all, it is part of their job to explore and develop new methods and technologies whereas practitioners normally only use these methods. This trait could lead to a systematically higher assessment of new methods in comparison to conventional methods.

To summarize the findings, we can state that on a basic level the results from both studies have a tendency towards the same results, but the results of study 1 could only be partly replicated due to the usage of a different sample. From a methodological viewpoint, the experience level of the sample probably influences the results. We argue that in future studies (especially for student and/or academic samples) prior experience should be analyzed more deeply and be discussed in line with the results.

B. Threats to Validity

Concerning the internal validity of the study, effects of maturation (e.g., fatigue) on the assessment of the two methods may be excluded because of the randomized order of the tasks and the coupled subjective rating. Standardized instructions were used, which minimizes influences caused by experimenter

expectancies towards the assessment of the two methods. This procedure additionally assures an identical treatment in each application of the study design.

With regard to the external validity of the study, we are aware that due to the small sample sizes (n=7 and n=11), the used example from avionics, performing the tasks via paper & pencil instead of PC supported modeling tools, and the ad-hoc selected samples, the results are only valid within a limited scope. The tasks, although small in size and complexity, were developed together with practitioners to ensure realism. Furthermore, participants were asked about the quality of the tutorial and if both approaches (FT and CFT) were understood sufficiently to apply them as well as the understandability of the provided materials and if the time was sufficient to work on the tasks. Some information regarding the participants' background (e.g., age, company affiliation time) was not allowed to be collected due to company restrictions.

In terms of conclusion validity, appropriate statistical test procedures were used and the calculated effect sizes are independent of sample size.

The laboratory conditions as well as the use of paper & pencil limit the ecological validity of the study. Construct validity is concerned with the quality of the operationalization of theoretical constructs into measurement design, sampling design, and design of the study [26]. One possible threat towards construct validity is the reactivity to the experimental situation. To reduce problems regarding this threat, we reduced experimenter interactions with the participants by including all relevant instructions, materials (tasks and corresponding questionnaires) in one package (the test booklet). In addition, the study procedure, material, and instrumentation were tested a priori in three pilot studies, yielding improvements concerning understandability, task descriptions, and procedure. Furthermore, anonymity as well as confidentiality was assured.

However, only a part of the system was used in order to avoid exceeding the given timeframe (i.e., 4 hours). Usually the scale of projects in the avionics domain easily reaches a size of thousands of system requirements and involves hundreds of engineers – a size that is difficult to handle in an experiment.

VIII. Conclusion and Future Work

Our work aimed at comparing two safety analysis methods (FT and CFT) by means of an empirical study. Therefore we first gave an overview of existing empirical studies in the field of safety engineering. Only a small number of studies are reported and to our knowledge, all of the studies used student samples and none of them analyzed FT or CFT or compared these methods. We then explained how the system design model, FT and CFT are related to each other. CFT adhere to the graphical logic of the system design model, whereas FT follows an independent graphical approach. Our research questions addressed two aspects: 1) the quality of modeling task solutions and 2) participants' subjective perception with regard to the methods. To answer these questions, we developed a comparative study design in a laboratory setting with realistic tasks and a realistic system model from the avionics domain.

With this experiment, we contribute to the body of knowledge by providing the first empirical study comparing FT with CFT and its replication with a practitioner sample. Based on the results of both studies and taking into consideration the threats to validity, we conclude that the participants rated CFT subjectively better than FT, although the use of CFT neither yielded significantly more correct solutions, nor did it lead to significantly more false solutions. One problem in comparing the task results for CFT and FT is the large number of missing solutions in the CFT condition for study 1 (4 out of 28) as well as for its replication (10 out of 44) compared to the FT condition (study 1: 0 missings; replication: 5 missings). One possible explanation is that FT was already used before the study and therefore it can be assumed that the participants already had a certain level of knowledge about FT, whereas CFT was unknown before. Consequently, some of the participants may have felt uncertain in applying CFT, thus preferring not to provide a false solution resulting in a higher number of missing solutions. Nevertheless, the number of correct solutions for CFT did not differ significantly from the number of correct solutions for FT. Taking into consideration the short duration of the tutorial (approx. 40 min) we interpret this result in favor of CFT. We hypothesize that with hands-on experience even better results for CFT can be achieved.

Based on the results, we conclude that CFT as a model-based approach can be beneficial for employees with little or no experience in fault tree analysis. However, if the employees have experience in FT, the implementation effort for introducing CFT in the company has to be balanced against a possible positive outcome regarding the employee's positive subjective assessment and the quality of the safety analysis models. Therefore, long-term analyses investigating the benefits of CFT in terms of time and costs are needed. It is also obvious that a tutorial of only 40 minutes cannot replace thorough training in a new methodology.

From a practitioner's viewpoint, the results of our work are promising because even with only a short tutorial, the participants were able to apply CFT with almost equally good results as FT. Also, it is possible to conduct an empirical study for manual parts of a model-based methodology. Further replications of the study design are needed to support the conclusions drawn from the results. These study designs should also include applying FT and CFT with PC supported modeling tools. Moreover, we plan to apply the study design to conduct studies with other, similar, model-based safety analysis methods, e.g., Static Event Fault Trees (SEFT) and Dynamic Fault Trees (DFT). Still, further research is needed to investigate whether CFT meets business goals (such as time, cost, and quality) when participants have a better knowledge in applying the method.

To address the question of scalability, we propose conducting a study in which the size of the system is varied. This will also answer the question of whether one technique is better for smaller or larger systems; e.g., does CFT enable safety engineers to handle the complexity of large systems?

ACKNOWLEDGMENT

We would like to thank the participants of our studies. We would also like to thank our colleagues Sonnhild Namingha, Bastian Zimmer and Anne Hess from Fraunhofer IESE for reviewing earlier versions of this paper.

REFERENCES

[1] SAE International, Guidelines for Development of Civil Aircraft and Systems. ARP4754A, 2010.

[2] Radio Technical Commission for Aeronautics Software, Considerations in Airborne Systems and Equipment Certification. DO-178B, 1991.

[3] Commission Regulation (EC), Official Journal of the European Union, No. 1702/2003 of 24 September 2003.

[4] European Aviation Safety Agency, "Certification Specifications and Acceptable Means of Compliance for Large Aeroplanes CS25", Amendment 12, 13 July 2012.

[5] International Electrotechnical Commission, Fault tree analysis (FTA), IEC 61025 ed2.0, 13 Dec 2006.

[6] C.A. Ericson, "Fault Tree Analysis - A History." Proceedings of 17th Intern. System Safety Conference Orlando, 1999.

[7] M.A. de Miguel, J.F. Briones, J.P. Silva, and A. Alonso, "Integration of safety analysis in model-driven software development". Software, IET, 2(3), 2008, pp. 260–280.

[8] W. Damm, A. Votintseva, A. Metzner, B. Josko, T. Peikenkamp, and E. Böde, "Boosting Re-use of Embedded Automotive Applications Through Rich Components", Elsevier's Electronic Notes in Theoretical Computer Science, Elsevier Science B.V., 2005.

[9] Y. Papadopoulos, and J.A. McDermid, "Hierarchically Performed Hazard Origin and Propagation Studies". Computer Safety, Reliability and Security. 1999.

[10] B. Kaiser, P. Liggesmeyer, and O. Mäckel, "A new component concept for fault trees". In SCS '03: Proceedings of the 8th Australian workshop on Safety critical systems and software, pp. 37–46, Darlinghurst, Australia. Australian Computer Society, Inc., 2003.

[11] R. Adler, D. Domis, K. Hoefig, S. Kemmann, T. Kuhn, J. Schwinn, and M. Trapp, „Integration of Component Fault Trees into the UML. Non-functional System Properties in Domain Specific Modeling Languages" (NFPinDSML'10), Workshop at ACM/IEEE 13th Intern. Conf. on Model Driven Engineering Languages and Systems, Oslo, Norway, 2010.

[12] D. Domis, K. Hoefig, M. Trapp, "A Consistency Check Algorithm for Component-based Refinements of Fault Trees". Proc. 21st IEEE Intern. Symposium on Software Reliability Engineering (ISSRE), San Jose CA, USA, pp. 171-180, 2010.

[13] T. Stålhane and G. Sindre, "A Comparison of Two Approaches to Safety Analysis Based on Use Cases". Lecture Notes in Computer Science, Vol. 4801, pp. 423-437, 2007. DOI: 10.1007/978-3-540-75563-0_29.

[14] T. Stålhane and G. Sindre "Safety Hazard Identification by Misuse Cases: Experimental Comparison of Text and Diagrams." Proc. of the 11th Int. Conf. on Model Driven Engineering Languages and Systems (MoDELS '08), Springer Lecture Notes in Computer Science Vol. 5301, pp. 721-735, 2008, DOI: 10.1007/978-3-540-87875-9_50.

[15] T. Stålhane, G. Sindre, and L. Du Bousquet, "Comparing safety analysis based on sequence diagrams and textual use cases." In Proc. of the 22nd Intern. Conf. on Advanced Inf. Systems Engineering (CAiSE'10), Barbara Pernici (Ed.). Springer-Verlag, Berlin, Heidelberg, pp. 165-179, 2010.

[16] T. Stålhane and G. Sindre, "Identifying Safety Hazards: An Experimental Comparison of System Diagrams and Textual Use Cases." Enterprise, Business-Process and Information Systems Modeling Lecture Notes in Business Information Processing, Vol. 113, pp. 378-392, 2012. DOI: 10.1007/978-3-642-31072-0_26.

[17] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology", MIS Quarterly 13(3), pp. 319–340, 1989.

[18] A.L. Opdahl and G. Sindre, "Experimental comparison of attack trees and misuse cases for security threat identification". Information and Software Technology, 51 (5), pp. 916-932, 2009.

[19] L. Briand, D. Falessi, S. Nejati, M. Sabetzadeh, and T. Yue, "Traceability and SysML Design Slices to Support Safety Inspections: A Controlled Experiment". Technical Report, Simula Research Laboratory, August 2010.

[20] G.J. Pai and J.B. Dugan: "Empirical Analysis of Software Fault Content and Fault Proneness Using Bayesian Methods". IEEE Trans. Software Eng. 33(10), 675-686, 2007.

[21] A. Paivio, Mind and Its Evolution: A Dual Coding Theoretical Approach, Lawrence Erlbaum: Mahwah, N.J., 2006.

[22] P. N. Johnson-Laird, Mental models. Towards a cognitive science of language, interference, and consciousness. Cambridge, England: Cambridge University Press, 1983.

[23] W. Schnotz, "On the relation between dual coding and mental models in graphics comprehension". Learning and Instruction, 3, pp. 247–249, 1993.

[24] W. R., Shadish, T. D. Cook, and D. T. Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin, 2002.

[25] R. Likert, "A Technique for the Measurement of Attitudes" Archives of Psychology, 140, pp. 1-55, 1932.

[26] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages". Psychometrika 12 (2), pp. 153–157, June 18, 1947. doi:10.1007/BF02295996

[27] F. Wilcoxon, "Individual comparisons by ranking methods". Biometrics Bulletin 1 (6), pp. 80–83, Dec 1945.

[28] L. Cronbach, "Coefficient alpha and the internal structure of tests," Psychometrika, Vol.16, 297-334, 1951.

[29] A. Field, Discovering statistics using SPSS. London, England: Sage Publications, 2009.