A Structured Hazard Analysis and Risk Assessment Method for Automotive Systems - A Descriptive Study

Kristian Beckers, Dominik Holling

Software Engineering
Technical University of Munich (TUM)
Boltzmanweg 3
Garching bei München
Germany
{*beckersk,holling*}*@in.tum.de*

Isabelle Côté, Denis Hatebur

ITESYS Institut für
technische Systeme GmbH
Emil-Figge-Str. 76
Dortmund
Germany
{*i.cote,d.hatebur*}*@itesys.de*

Abstract

The 2011 release of the first version of the ISO 26262 standard for automotive systems demand the elicitation of safety goals following a rigorous method for hazard and risk analysis. Companies are struggling with the adoption of the standard due to ambiguities, documentation demands and the alignment of the standards demands to existing processes. We previously proposed a structured engineering method to deal with these problems developed in applying action research together with an OEM. In this work, we evaluate how applicable the method is for junior automotive software engineers by a descriptive study. We provided the method to 8 members of the master course *Automotive Software Engineering (ASE)* at the Technical University Munich. The participants have each been working in the automotive industry for 1 to 4 years in parallel to their studies. We investigated their application of our method to an electronic steering column lock system. The participants applied our method in a first round alone and afterwards discussed their results in groups. Our data analysis revealed that the participants could apply the method successfully and the hazard analysis and risk assessment achieved a high precision and productivity. Moreover, the precision could be improved significantly during group discussions.

Keywords: requirements, ISO 26262, automotive, safety, structured method, empirical study

1. Introduction

The safe construction and development of road vehicles is a very complex task, due to the ever increasing features for drivers, e.g., adaptive cruise control. Considering all relevant aspects of safety is of paramount importance during the product development. With the release of ISO 26262 - Road vehicles for Functional safety in November 2011, the automotive sector benefitted from a consistent functional safety process for developing and constructing road vehicles [1]. The standard's scope covers electronic and electric (E/E) systems for vehicles in series production with a max gross weight up to 3500 kg. Since ISO 26262 is a risk-based functional safety standard addressing malfunctions, its process makes use of a hazard analysis to determine the necessary risk reduction to achieve an acceptable level of risk. The necessary risk reduction is described by an automotive safety integrity level (ASIL). Following an ISO 26262-based process guides through the elicitation of safety goals, safety requirements up to the level of technical safety requirements to reduce risks. This includes components/systems inside as well as outside the system boundary. Currently, all vehicle manufacturer (OEM) and their suppliers working on adapting their processes to be compliant to ISO 26262.

The standard adaption is difficult, because of the ambiguous and lengthy texts in the standard. We proposed a structured and step-wise method for the hazard and risk analysis parts of the standard that is easy to follow [2]. The work was developed in action research in collaboration between FORD Germany, University Duisburg-Essen, and ITESYS. However, the work was not empirically evaluated with an alternative sample of engineers. We contribute such an evaluation in this paper. In particular, we evaluated the following concerns:

- A measure of true positives and *precision* (true positives / all values) of the application of our method steps in order to determine how well junior engineers can apply these steps
- A measure of difference in *precision* between individual assessments and group assessments
- A measure of difference in *productivity* (filled relevant parts in a template) between individual applications of our method

The Technical University of Munich (TUM) has a dedicated Master's program for Automotive Software Engineering. The selected students educated in this course benefit from lecturers from the automotive domain. We selected a course for functional safety based on the ISO 26262 as the context for our study. We provided a template based version of our method proposed in [2]. We decided to use a template-based version of our method, due to time constraints of the study. We plan to redo the study in the future with an introduction into UML, our profile, and our tool support. Note that we use UML in our method presented in [2], because UML is a fairly well-known modelling notation that can be easily extended with stereotypes in profiles. We created such a profile

for the ISO 26262 terminology. Our profile can be mapped into SysML or other refined UML notations with little effort. We did not want to prescribe a specific notation such as SysML, because we want our work to be applicable in a larger scope. In our course at TUM 8 participants took part in our study. These all had at least 1 year of experience working as a student in the automotive domain. Moreover, we assessed their knowledge in software and safety engineering with a questionnaire. The result shows (see Sect. 4) that the sample represents junior engineers in the automotive with on average rather limited experience in hazard analysis and risk assessment.

Our results show that the method is applicable for our intended target audience: junior engineers in the automotive industry. Nevertheless, we found several points for improvement that our study revealed. Finally, we show that the precision of our method improves when applying it during group sessions. This result confirms the importance of group discussions during hazard analysis and risk assessment.

The remainder of this paper is structured as follows. Section 2 presents background to our research and Sect. 3 contains related work. Section 4 shows the planning phase, Sect. 5 details our operational phase, Sect. 6 contains the data analysis, and Sect. 7 shows the results of our exit questionnaires and interviews. We report on threats to validity of our study in Sect. 8. Section 9 concludes and provides directions for future research.

2. Background

2.1. ISO 26262

ISO 26262 [1] was derived from the generic functional safety standard IEC 61508 [3]. It is aligned with the automotive safety life-cycle including specification, design, implementation, integration, verification, validation, configuration, production, operation, service, decommissioning, and management. ISO 26262 provides an automotive-specific risk-based approach for determining risk classes that describe the necessary risk reduction for achieving an acceptable residual risk, called *automotive safety integrity level (ASIL)*. The possible ASILs are *QM*, *ASIL A*, *ASIL B*, *ASIL C*, and *ASIL D*. The ASIL requiring the highest risk reduction is called ASIL D. For functions with ASIL A, ASIL B, or ASIL C, fewer requirements on the development processes, safety mechanisms, and evidences are given in ISO 26262. In case of a QM rating, the normal quality measures applied in the automotive industry are sufficient.

2.2. Our Structured Method for Hazard and Risk Analysis

We proposed a structured method based on UML models supported by a tool in order to derive technical safety requirements [2]. We used an item definition and other engineering documents created using a ISO 26262-compliant process as an input to develop a technical safety requirements documentation. According to ISO 26262, the item is a set of functions realised by the system to be built. We describe the method in more detail in the description of our study setup (see Sect. 4).

3. Related Work

We are not aware of any publications about a structured and safety requirements analysis approach compliant to ISO 26262 including Hazard and Risk Analysis that has been empirically evaluated.

The following holistic approaches to safety management for the automotive domain have been proposed. Tang et al. [4] present a holistic approach for concerning the product lifecycle of automotive development. The entire safety lifecycle including safety requirements analysis is presented by Baumgart [5]. The Safety Management System and Safety Culture Working Group provides guidance on functional safety development by different means, e.g., brainstorming, HAZOP, checklists, FMEA [6]. Jesty et al. [7] give a guideline for the safety analysis of vehicle-based systems, including system analysis, hazard identification, hazard analysis, identification of safety integrity levels, FMEA, and fault tree analysis. All of these works lack an empirical evaluation.

Researchers have proposed the following methods particular for hazard analysis in the automotive domain. Mehrpouyan [8] proposes a model-based hazard analysis procedure (based on SysML models) for the early identification of potential safety issues caused by unexpected environmental factors and subsystem interactions within a complex safety-critical system. The proposed methodology additionally maps hazard and vulnerability modes to specific components in the designed system and analyses the elicited safety requirements. Giese et al. [9] present an approach that supports the compositional hazard analysis of UML models described by restricted component and deployment diagrams. Papadopoulos and Grante [10] propose a process that addresses both cost and safety concerns and maximises the potential for automation to address the problem of increasing technological complexity. It combines automated hazard analysis with optimisation techniques. None of these approaches has been empirically evaluated, as well.

Empirical case studies exist for *Hazard Analysis regarding fault trees*. None of the existing studies evaluates a structured and standard compliant method to ISO 26262.

Mouaffo et al. [11] conducted a controlled experiment to compare two techniques for modelling stochastic dependencies of events in safety-critical systems. Both techniques are based on fault-trees. The authors focused on the applicability of the method from the viewpoint of the users and their evaluation focused on criteria such as the simplification of the notations, tools support and knowledge base available. Martins et al. [12] propose novel ways of annotating fault trees for an easier derivation of functional requirements. Jung et al. [13] conducted an experiment on comparing Fault Trees Analysis and those of Component Fault Trees Analysis. The experiment was first run with university researchers and later on replicated with practitioners. Stålhane et al. [14] analyses the Failure Mode and Effect Analysis (FMEA) technique in comparison to Misuse Cases regarding perceived ease of use, while Venkatesh et al. [15] analysed the perceived usefulness by means of the Technology Acceptance Model. In addition, Stålhane et al. compared Misuse Cases to use-case diagrams and textual representations of use cases [16] and System Sequence Diagrams [17].

In *security engineering* controlled experiments have been done to measure the effectiveness and efficiency of e.g. vulnerability analysis techniques [18] security patterns [19] and the application of standards in the aviation domain [20], as well as the effectiveness of hierarchies for security requirements patterns [21]. Our study design is inspired by these studies. Note that we only take elements of these studies, since we deliver at this point only a descriptive study and not a fully fledged controlled experiment.

Chen et. al. [22] describe the use of the EAST-ADL2 architecture description language to integrate safety analysis techniques, a method for developing and managing Safety Cases, and a systematic approach to model-based engineering. This paper does not focus on hazard analysis and also does not describe a empirical study.

4. Planning the Study

Our structure for planning and reporting on the study is inspired by the work of Scandariato et al. [18].

4.1. Goals

We define the goals of this study by using the template defined by Basili et al. [23] in the following.

Purpose The purpose of the study is to assess the effect of

Object of Study automotive hazard analysis and safety goal elicitation compliant to ISO 26262

Focus on both individual and groups productivity and quality of the analysis results

Stakeholder from the point of view of junior safety analyst in the automotive domain

Context and in the context of a master's level lecture for Safety Engineering as part of the course Automotive Software Engineering at the Technical University of Munich (TUM).

4.2. Participants

We surveyed the background of our participants by administering a questionnaire at the beginning of the study. The participants are 8 students at a lecture for safety engineering in the automotive domain. All of the students are enrolled in the Master's level course Automotive Software Engineering at the Technical University of Munich (TUM). The course is positioned in the second year of a 2 years master's program (see Fig. 1).

All of the participants are working students in the automotive domain with up to 4 years of working experience (see Fig. 2). Their working areas are mainly

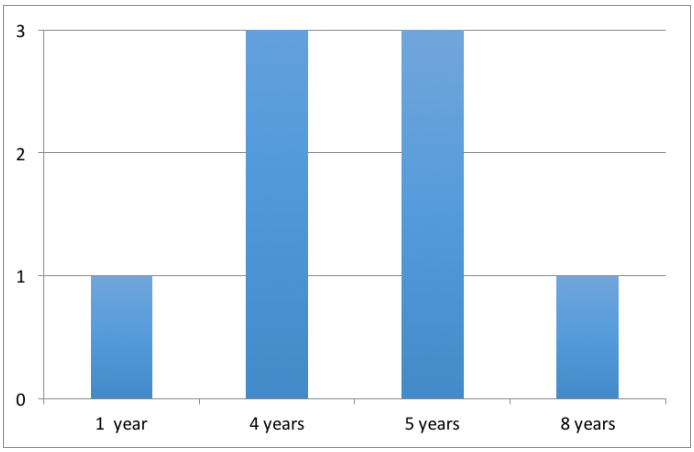

Figure 1: Study Time on a University

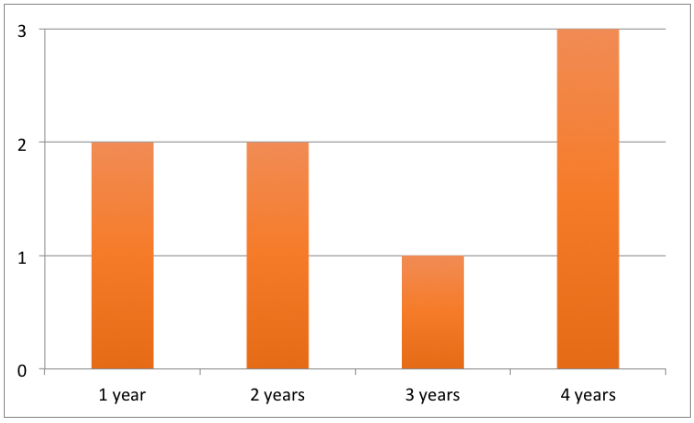

Figure 2: Work Experience

concerning software implementation (see Fig. 3). Note that students sometimes work in more than one area.

We also investigated their skill level related to hazard analysis. We asked them in a questionnaire how they would rate their skill level and afterwards provided a multiple choice question for that area. These questions were independently reviewed by 2 senior researchers. We rated the quality of the possible answers to match the skill levels. The answer options were all correct, but some answers had ambiguities in them. We used the well-established scale of expertise from Ernaut [24] to classify the skill level of the participants. We present in the following figures two bars for each element in the scale. The left bar represents the self-assessment of the students and the right bar represents assessment based on our questions, e.g., a student had to pick the least ambigues answer to be rated as an expert.

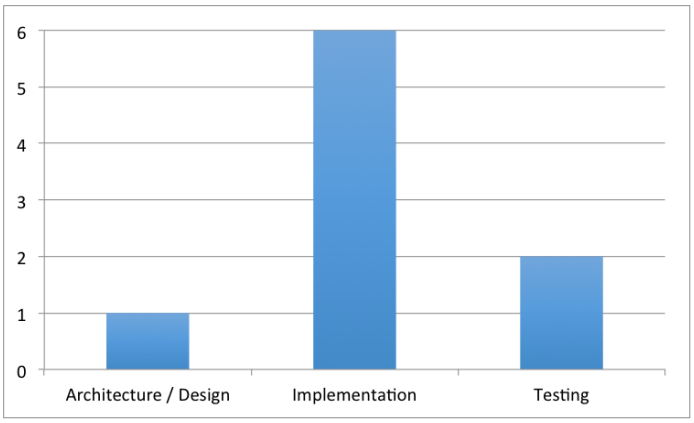

Figure 3: Working Areas in Software Engineering

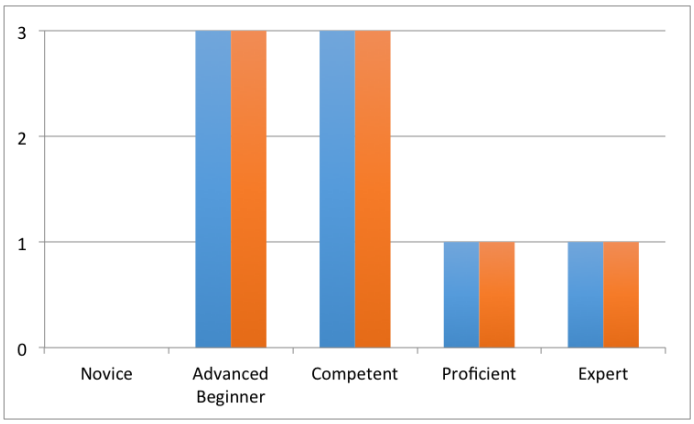

Figure 4: Experience in Automotive Bussystems

We measured their knowledge with regard to important systems in the automotive domain. Their experience in Automotive Bussystems (see Fig. 4), e.g. CAN Bus [25] and Realtime systems (see Fig. 5) were mostly in the skill levels of advanced beginner and competent. We analysed their skill level in software engineering, as well. In software engineering most of the participants are advanced beginners (see Fig. 6).

We measured the expertise in Safety Engineering and most of our participants have the skill level competent (see Fig. 7). Their knowledge in hazard analysis is almost evenly distributed between the levels of advanced beginner and competent (see Fig. 8), which is what we expect from a junior safety engineer. However, most of the participants had no experience with the ISO 26262 compliant risk assessment (see Fig. 9).

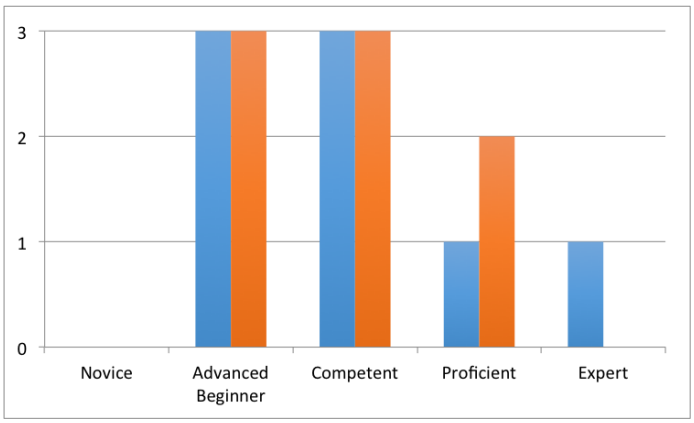

Figure 5: Experience in Realtime Systems

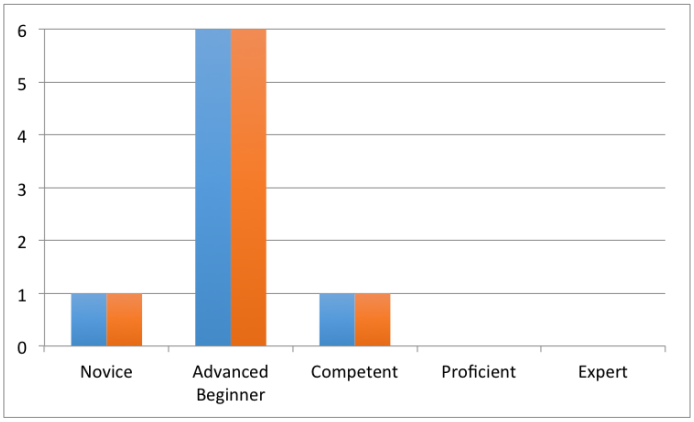

Figure 6: Experience in Software Engineering

195 It is worth mentioning that almost all self-assessments of our participants matched the results of the questions, except for two difference in the realtime systems (see Fig. 5) and hazard analysis (see Fig. 8) were students claimed to be experts but were rated only as proficients according to the answer to our questions.

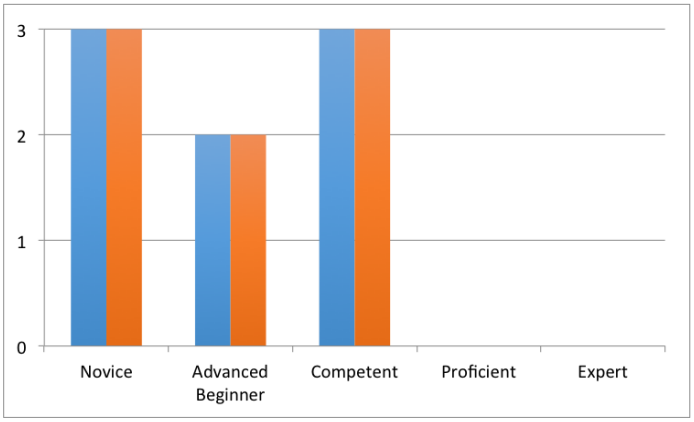

Figure 7: Experience in Safety Engineering

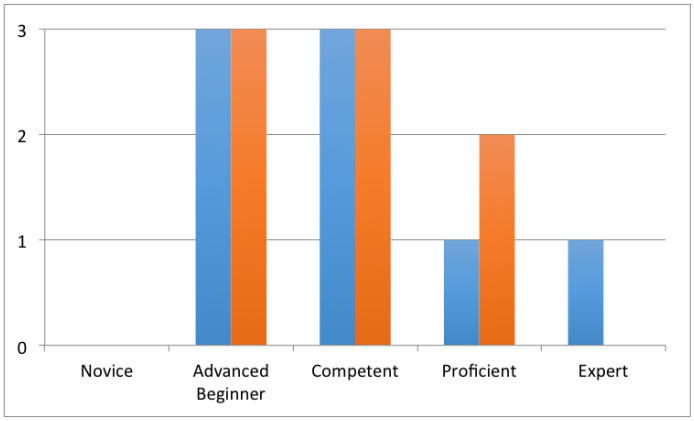

Figure 8: Experience in Hazard Analysis

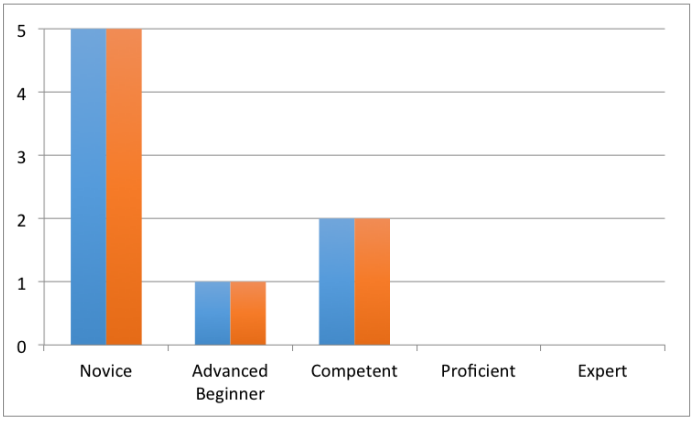

Figure 9: Experience in ISO 26262 Risk Assessment

4.3. Study Objects

For this study we needed to select a structured method that contains all the relevant steps for conducting a hazard analysis and risk assessment for the automotive domain compliant to the ISO 26262. We based this on our previous work with FORD that fulfils this criteria [2]. We refer to the related work of that publication that no equally structured publicly available work exist and we checked for new publications before running the study.

The participants should apply our method to an example that is small in size, so we could keep the entire study within the time constraints of 2,5 hours. However, the example had to be rich enough so that different hazards and risks can emerge. We selected the example of an electronic steering column lock system (ESCL) as introduced in [2], because it fits these criteria.

The main function of the ESCL is to provide lock and unlock commands to the lock actuator automatically to enhance theft protection for vehicles with a power button instead of a standard key. The ESCL interacts with the following elements in the environment, namely the the driver, the lock actuator, and the vehicle. The item (the ESCL) controls the lock and the unlock commands, and the lock actuator observes these information and reacts accordingly. The driver presses the power button to crank or stop the engine. The vehicle moves at a certain speed and therefore controls the phenomenon Speed. The following requirements shall be considered for the ESCL: *R1 - The steering column shall be locked, when the driver wants to immobilise the vehicle* and *R2 - The steering column shall be unlocked, when the driver wants to drive*. The requirements both contain the lock actuator, while only referring to driver and vehicle.

4.4. Tasks

We proposed a method [2] for conducting a hazard analysis and risk assessment according to ISO 26262. The aim of the analysis is to identify and classify the potential hazards of the item and to formulate safety goals related to the prevention or mitigation of these hazards in order to achieve an acceptable residual risk. ISO 26262 demands a definition of the item, its basic functionality, and its environment. We provided this description to the students in the form of a graphical model as described in [2]. In contrast to this previous work, we provided textual templates for the students to fill out instead of graphical models to instantiate. The reason for this change is the increased time and education the students would need to draw the models. The students had to complete the following tasks (based on the steps of our method).

Task 1. Instantiate Fault-Type Guide-Words. We proposed a set of so-called *fault-type guide-words* inspired by the HAZOP standard [26]. The guide-words help the developer to consider all relevant faults. We provided the following guide words for the students during the study: *no, unintended, early, late*. Note that due to the time constraints we use only 4 of 8 keywords prescribed by the HAZOP standard. Each guide-word has to be instantiated for the functions specified in the provided *item definition*. For each requirement, all these fault-types are checked if they have to be considered, are not possible, or are covered by other faults. For each considered fault, the students have to

describe the effect on the system level and not on component or vehicle level. On the system level, the elements of the item are visible, e.g., actuators. In vehicle level descriptions, only phenomena that can be observed or controlled by the driver or other persons are used. The component level contains descriptions of internal interfaces, e.g., CAN [25] messages. For faults rated not to be considered, either a description why it is not relevant or a reference to at least one other fault specifies that it covers this fault as well.

Task 2. Situation Classification. We provided a list of driving situations to the students, namely *driving-at-high-Speed*, *driving-at-low-Speed*, *parking-Maneuver-Situation*, *standstill-Engine-Off-Situation*, *stand-still-Engine-On-Situation*, and *to-tow-away*. Using this list, the participants had to rate if a situation is relevant for the described item with its requirements or not. Students are allowed to create hierarchies of driving situations and if a more abstract situation is rated, it is not necessary to rate the special situations. This is done to reduce the overall effort of the hazard analysis, because the special situations are not considered in the following steps. If a situation is rated as not being relevant, either a reference to another situation that includes this situation is given, or a rationale has to be provided.

Task 3. Hazard Identification. For each fault/function combination, all situations that could lead to a potential hazard had to be identified by the participants in the list of situations being relevant. The participants have to describe the effect on the vehicle level, i.e., what behaviour could occur in case of a potential item's malfunction. Based on the effect on the vehicle level, they had to describe the hazards and possible consequences. Hazards are defined in terms of the conditions or events that can be observed at the vehicle level (e.g, by the driver). A hazard is $\ll$*caused by*$\gg$ a set of faults and refers to situations $\ll$*when*$\gg$ it can occur. These refer back to the information elicited in Tasks 1 and 2. It is important that each relevant situation is referenced by at least one hazard and each of the faults has to be considered by at least one hazard.

Task 4. Hazard Classification by Severity, Exposure, and Controllability. The objective of the hazard classification is to assess the level of risk reduction required for the hazards. To classify the hazard, the following steps need to be performed by the participants. They had to estimate the potential severity, exposure, and controllability using the predefined classes in ISO 26262 e.g. S0 (no injuries) and S1 (light and moderate injuries). The participants had to provide a rationale for each class selected. We provided the descriptions of and examples for the classes from the standard [1, Part 3, Appendix B]. Based on these estimations, the ASIL is determined automatically according to the corresponding ISO 26262 table. For example, a rating of S3, E4, and C3 leads to ASIL D. This table was provided to the participants, as well.

Task 5. Define Safety Goals. Safety goals have the attributes ASIL, safe state, and fault tolerance time. The ASIL is a measure of necessary risk reduction. The safe state is a state that shall be entered to avoid a hazard. The fault tolerance time is the time an actuator state can be unsafe before the situation

becomes hazardous, e.g., an undue brake intervention may have a fault tolerance time of 100 ms in certain situations. ISO 26262 requires that at least one safety goal is assigned to each hazard rated as ASIL A, B, C or D. It is not necessary to define safety goals for hazards rated as QM.

One safety goal can address several hazards. A hazard can be addressed by more than one safety goal. ISO 26262 requires that if a safety goal can be achieved by transitioning to or by maintaining one or more safe states, then the corresponding safe states are specified.

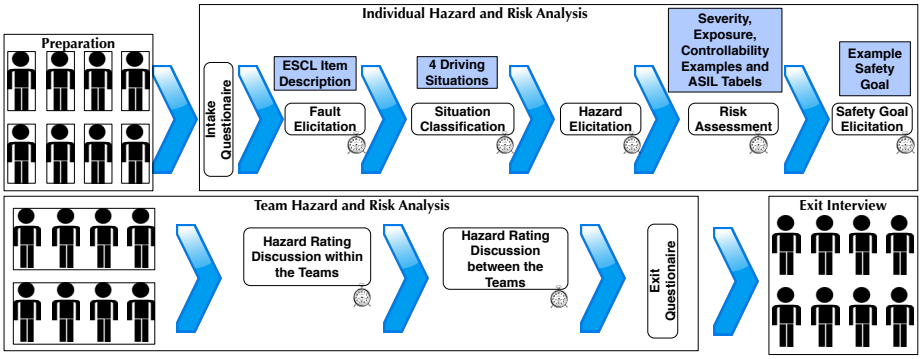

Figure 10: Design of the Study

4.5. Design of the Study

We illustrate our study in Fig. 10, which consists of two lab sessions. In the first lab session 8 participants followed our procedure for ISO 26262 hazard and risk analysis for an electronic steering column lock. The second lab session the participants were split into 2 teams of 4 members each. They discussed their results in teams with the goal of finding the most severe hazard with the highest risks. This is essential, because hazard analysis often focus on the hazards with the highest risk. If these cannot be reduced to acceptable risk levels the entire item may not be included into the vehicle. Moreover, the participants should agree on all the information leading to the elicitation of that hazard: faults and situations, as well as agree on controllability, severity, and exposure classes and the subsequent ASIL. Finally, they should state one safety goal. After the discussion within the teams finished the next assignment started. The two teams had to agree on the most sever hazard including its ASIL values and resulting safety goal.

In summary, we decided for a design concerning one treatment applied individually and in groups. We randomised the assignments of participants to groups. The same object of study was chosen for individual and team assignments, because the previously learned information about the hazard should be discussed in groups. We observed the group discussions and evaluated if the participants had enough information available from the application of our method to engage in structured arguments. These are essential for finding the

hazards with the highest risks in teams. We assessed the knowledge of participants of automotive systems, hazard analysis, and software engineering with a questionnaire that include self assessments and its validation via multiple choice questions. We used the Ernaut [24] scale for knowledge acquisition and the multiple choice answers contained varying levels of granularity and we expected for people that claimed to be expert to find the answer with the least granularity and for proficients the answer with the least ambiguity or the 2nd least and so forth.

After the group study, we asked the participants for a self assessments for their knowledge about functional safety and hazard analysis with the same questions as in the initial survey. We asked them for the time they required for the processing of the questions individually and in groups. Moreover, we inquired if they learned hazard analysis via application of the method. We also asked if they learned more in the individual or group discussions and if they are convinces that the proposed ASIL and safety goal is correct. Finally, we asked all the participants in a group for feedback with regard to positive and negative aspects of our method.

4.6. Hypothesis

Our goals for the study described in the beginning of this section are to measure the quality of the results and the productivity of junior safety analysts. We characterise the quality of the results primarily by the number of correct results (faults, driving situations, hazards and safety goals), because in particular the elicitation of relevant hazards found mean a more complete safety analysis, and subsequently safer cars. In addition, the number of errors (missing or false attributes of faults, driving situations, hazards, and safety goals) are relevant, because they result in a waste of resources of the analysis team and the teams for the ISO 26262 reviews.

Correct results are called true positives (TP) and the errors are called false positives (FP). The number of fields a participant can instantiated in our templates (RH). The productivity (PRO) is quantified considering the amount of instantiated fields, true positives and false positives. Note that Scandariato [27, 18] defines productivity as tasks processed per time unit. We define this notion differently from the definition of Scandariato [27, 18], because we are interested in the productivity within a given time frame with respect to the entire hazard and risk analysis.

The precision (PRE) is quantified as ratio of correct instantiated fields in our templates over the total amount of fields in all templates used in our study. We consider the number of errors in the precision, which scales with respect to the total amount of results. This corresponds to the reasonable assumptions that, when more work is done, more mistakes are made. Note that Scandariato [27, 18] defines precision as the TP with respect to TP and FP. We define this notion differently from the definition of Scandariato [27, 18], because we are interested in the precision within a given time frame with respect to the entire hazard and risk analysis.

In addition, we are interested on if our junior safety analysts could process our method steps and the templates representing them in the time given. We

measure productivity in this sense as a simple indicator for task being manageable in the timeframe given. The number of relevant instantiations (RI) is the number of fields instantiated that should be instantiated. This considers only that a value is contained in a field that should contain a value. Note that it does not consider if this is the correct value. The productivity (PRO) is quantified as ratio of relevant instantiated fields in our templates over the total amount of fields. This value serves as an indicator for the comprehensibility of our templates. It reveals that if relevant fields are not even instantiated by our participants, these are cause for concern and should be revised.

Table 1: Terminology

ID	**Measure**	**Definition**	**Formula**
TP	True Positive	A field in a template is correctly instantiated (correct result that experts found previously or verified)	
FP	False Positive	A field in a template is not correctly instantiated (e.g. incorrect fault)	
RH	Number Fields	The total number of fields that can be instantiated by a participant	
PRE	Precision	Percentage of correctly instantiated fields by the participants	TP / RH
RI	True Instantiation	Number of relevant fields instantiated	
PRO	Productivity	Percentage of instantiated fields	RI / RH

Using the above definitions, we report on *precision*, as well as *productivity* of our participants. Firstly, we report on precision individually when applying the entire method. Secondly, we report on the comparison of *precision* of individual and groups only on the elicitation of hazards. Hazard are the essential output of our method and deserve special consideration. Thirdly, we report on the *productivity* of our method.

We are interested in the *precision* of the results of our students. As a hypothesis we are interested in knowing whether, on average, the number of correct results is pre-dominant (at least 80 %) regarding all instantiated templates including false positives. Admittedly, the choice of a 80% threshold is somewhat arbitrary, it has been chosen before by Scandariato et al. [27] for measure the precision of the Microsoft STRIDE threat analysis technique. The authors stated that this number is often regarded as a valid reference for precision in domains such as information retrieval. We formulate our null hypothesis as follows.

$$H_0^{PRE}\text{: } \mu \{ \text{ PRE } \} \leq 0.80$$

We are in particular interested in the *precision* of the description of hazards including the ASIL determination and safety goals of individuals in comparison to the subsequent work in *teams*. Our null hypothesis here is that individuals produce the same percentage of correct hazards and safety goal descriptions as the teams.

$$H_0^{PRET}\text{: } \mu \{ PRE_{INDIVIDUAL} \} = \mu \{ PRE_{TEAM} \}$$

Moreover, we are interested in the *productivity* with regard to how many relevant fields are instantiated. If relevant fields are missing the overall applicability of our method is in jeopardy, because missing information in any hazard analysis results to overlooked details and subsequently possibly even missed hazards. As a hypothesis we are interested in knowing whether, on average, the number of relevant instantiated fields is pre-dominant (at least 90 %) regarding all fields in all templates. We decided to aim for a higher value as the value demanded for precision, because we do not consider the correct values of the instantiation and check simply that relevant fields are instantiated. We formulate our null hypothesis as follows.

$$H_0^{PRO}\text{: } \mu \{ \text{PRO} \} \leq 0.90$$

We use our results of the application of the method presented in [2] as a baseline for the values of true positives in this study and also the time frame given to the participants of the method is based on our previous experiences.

5. Operation of the Study

5.1. Training of the Participants

The study is embedded into a master program on automotive software engineering [28] and in particular in a course on safety engineering. The course is held by a practitioner that consults with major automotive OEMs and is taught as a compact course within a week of 8 hours daily. The first 4 hours of each day are a lecture, while the remaining hours are dedicated to labs.

On the first day of the course the students received in the lecture part a introduction in the development of the ISO 26262 and its legal implications. This introduction was held by a recently retired lawyer that worked with the standard for a leading automotive manufacturer. Different to the security domain, it is not sufficient in the safety domain to convince a certification body and get a certificate for standard compliance (e.g., for ISO 27001 compliance). In contrast, engineers follow the process outlined in the standard and document a so-called *safety case*. These documents are the basis of a legal defence if an accident happens with an automobile. The safety case has to show that the engineers did their due diligence in particular with regard to hazard and risk analysis and followed all best practices. Finally, they have been told that they have to sign with their name that they are convinced the hazard and risk analysis is correct. We included such a question in our exit questionnaire (see Sect. 7) for the students to find out if they would sign the analysis with their name and hence trust in the correctness of the results.

Our study was conducted on the fourth day of the course. Each participant received a basic training on hazard analysis in the morning, which did not contain our specific method. We taught that during the beginning of the afternoon. The participants received in the lecture in the morning a detailed introduction into working with the ISO26262 including the demands for hazard and risk analysis. The lecturer used a different example than the ESCL, which we used in our study. This lecture mimics an introduction for automotive engineers to ISO 26262 compliant hazard and risk analysis.

5.2. Execution of the Study

We initiated the lab on the fourth day of the course with an introductory talk about our hazard and risk analysis method that was developed in collaboration with FORD, ITESYS, and the University Duisburg-Essen (UDE) in 2013 [2]. The item description was provided in the form of a UML context diagram using the previously introduced notation (see Sect. 2). Moreover, we explained the benefit of a structured method for this effort and our interest in how well the students can apply the method. The introduction lasted 20 minutes. We also brought a physical ESCL Lock for the participants so they could touch and inspect the part in reality next to having seen our model. The entire lab session lasted for 3 hours and we started at 2 PM.

We provided an entry questionnaire to the participants to assess their knowledge in automotive hazard analysis, software engineering, and ISO 26262 (see Sect. 4 for details). The students worked on answering the questionnaire for 20 minutes in the lab.

Afterwards, we handed the students the exercise sheets out on printed paper. We instructed them to work on the sheets individually and direct all questions to us and not to their fellow students. We asked them not to use laptops and other digital devices. The participants worked on their paper sheets for 90 minutes and applied the steps of our method.

As a next step, we had a 20 minute session in which the participants were split into 2 groups of 4 participants each. We asked them, based on the information gathered in the exercise, to identify the most severe hazard and agree on its attributes. We made this choice because the limited time did not allow us to discuss all hazards and exercise sheets in groups. Our prime interest was to see if the teams would agree on one hazard that deserves the most attention during the safety analysis. Moreover, we are interested to know if the students would be able to agree on its risk values, in particular given the time constraint. Note that risk values are the values for ASIL and the severity, controllability and exposure that lead to the ASIL determination. Finally, we asked the participants to discuss all together given the same task. We provided a time window of 10 minutes for this discussion.

The study terminated with an exit questionnaire that the participants had to fill out in the remaining time. The questionnaire asked if they had learned about ISO 26262 hazard and risk analysis, and assessed the different experience of working on the hazard analysis alone and in groups (see Sect. 7 for further details). Note that the entry and exit questionnaires are provided in German.

However, two senior researchers that are native German speakers translated the results to English.

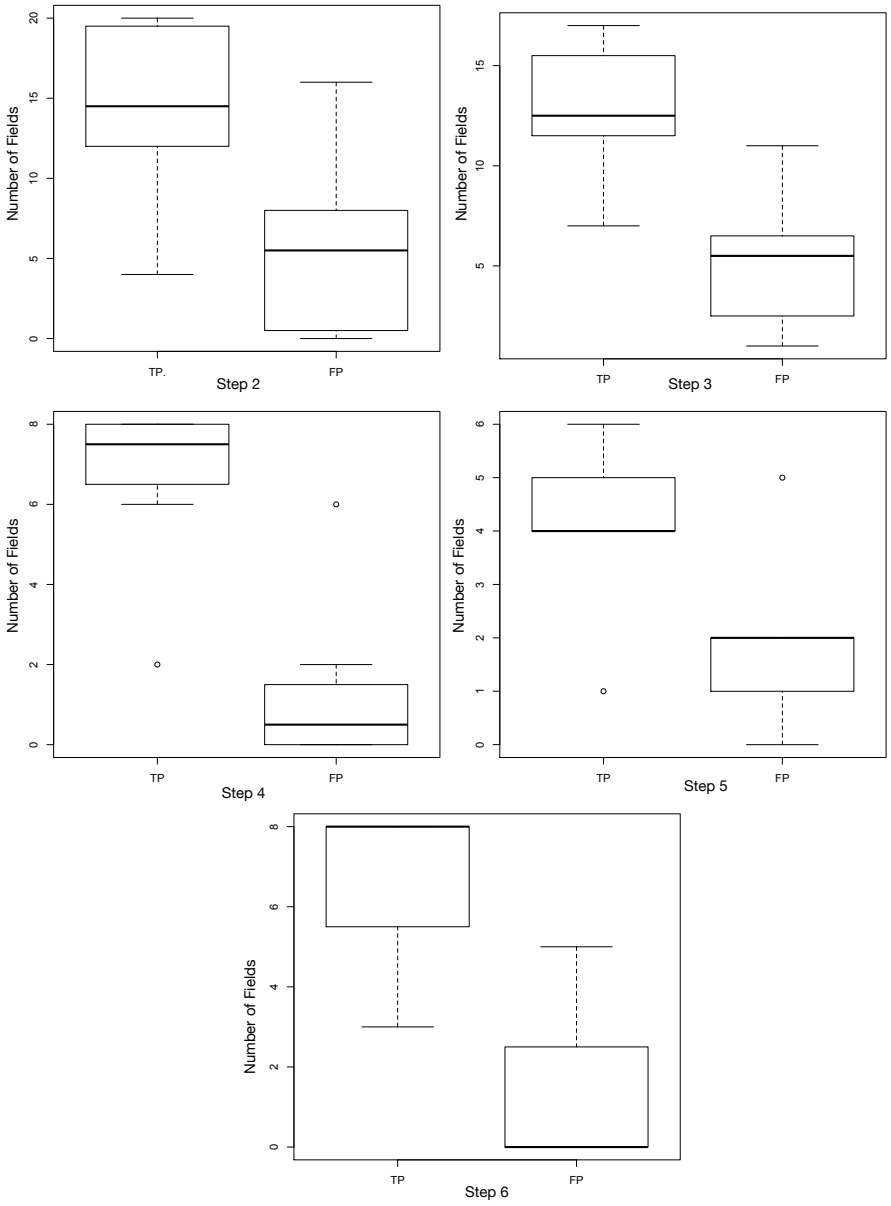

Figure 11: Overview TP/FP Method Steps

5.3. Measurement Procedure

All questionnaires and exercises were done using paper sheets. The filled-out sheets of the study were returned immediately after the study. The filled-out sheet were validated and digitalised by a senior researcher. The results of each student were compared against the results, we documented when we developed and applied the method previously (c.f. [2]). In case the answers matched with the previous research, we evaluated them as a TP. If the the answers did not match, we had a senior safety engineer evaluate the answer and evaluated them as a FP. For the case that the expert viewed the result as a TP that is different from the initial result, we report on these cases in the analysis section (see Sect. 6).

The students turned in the entry questionnaire once they filled them out. The questionnaire is evaluated by a senior researcher and the results were reported within the planning section of our study (see Sect. 4). The report included the self-assessment and the objective assessment with our multiple choice answers. The questions itself were designed by one senior researcher and validated by another.

We did not use time tracking of the individual steps of the method of each participant. We provided a reminder 10 and 5 minutes before the time limits were reached. The participants managed to stay within the provided time limits and some participants finished even earlier.

The exit questionnaire was distributed after the exercise was finished and collected immediately afterwards. The sheets were checked by a senior researcher and its answers checked. We asked the participants for another self-assessment in the same fashion as in the entry questionnaire. The senior researcher compared the answers and also reported on the participants assessment of the method. We report on the findings in Sect. 7.

6. Data Analysis

This section provides some descriptive statistics of the application of our method and shows the evaluation of our hypothesis. The material used in our study is provided online[1]. We analyse our data with the statistics software R[2] and report on our results in the following.

6.1. Precision Individuals: H_0^{PRE}

Table 2 describes the results of our study. We present the scores of the 8 participants in the first lines of the tables. The participants received 1 point for every field they instantiated correctly. The results of our method with the experts in the previous work [2] is presented in the line MAX. These are the maximal number of points a participant could achieve in our study and the baseline for our study. We describe the mean, standard deviation and also contains the 95% confidence interval for all our method steps (one-sample Wilcoxon

[1] `https://sites.google.com/site/researchhara/`
[2] `https://www.r-project.org`

test). We also illustrate these results in boxplots in Fig. 11. We report on our findings per step (see Sect. 4.4) in the following paragraphs.

Table 2: Evaluation of our Method: Precision

Participant	Step2	Step3	Step4	Step5	Step6
1	10	11	8	8	4
2	20	15	7	8	4
3	15	13	2	8	5
4	20	17	8	7	6
5	4	7	8	4	4
6	14	12	7	3	5
7	19	12	6	8	4
8	14	16	8	8	1
MAX	20	18	8	8	6
μ	14.5	12.88	6.75	6.75	4.13
σ	5.5	3.18	2.05	2.05	1.45
CI	[9; 19.5]	[9.5;16]	[5;8]	[5.5;8]	[2.5;5]
Precision	**0.725**	**0.715**	**0.843**	**0.843**	**0.687**

Step 2. With regard to our hypothesis, we can state that Step 2 is barely below the targeted value of 0.80 with 0.725 (see Tab. 2). The confidential interval (95% interval with one sample Wilcoxon Test) is [9:19.5] and the highest in our test and reflects our problems. The p-value is 0.03552 (Wilcoxon signed rank test). These values reflect the difficulty that the participants had with eliciting the faults in our study. The combination of function names with the fault type guide words worked well, but the building of hierarchies was a significant problem for our participants. This leads us to aiming for improvement for this step. Hence, Step 2 needs to be improved, in particular with regard to the hierarchies.

Step 3. With regard to our hypothesis, we can state that Step 3 is barely below the targeted value of 0.80 with 0.715 (see Tab. 2), as well. The confidence interval (95% interval with one sample Wilcoxon Test) is [9.5:16] is lower than in Step 2, but still higher than in the following steps. The p-value is 0.01415 (Wilcoxon signed rank test). These values reflect the difficulty that the participants had with ranking the driving situations in our study. Hence, Step 3 needs to be improved with more guidance for ranking driving situations.

Step 4. With regard to our hypothesis, we can state that Step 4 is above the targeted value of 0.80 with 0.843 (see Tab. 2), as well. The participants managed to execute this step reasonably well. The confidence interval (95% interval with one sample Wilcoxon Test) is [5:8] lower than the previous steps. The p-value is 0.09751 (Wilcoxon signed rank test). We are confident that Step 4 was sufficiently executed.

Step 5. With regard to our hypothesis, we can state that Step 5 is above the targeted value of 0.80 with 0.843 (see Tab. 2), as well. The confidence interval (95% intervall with one sample Wilcoxon Test) is [5.5:8] lower than the previous steps. The p-value is 0.1814 (Wilcoxon signed rank test). The participants managed to execute this step reasonably well. We have seen only minor problems with setting the severity, controllability, and exposure variables. We are confident that Step 5 was sufficiently understood.

Step 6. With regard to our hypothesis, we can state that Step 6 is below the targeted value of 0.80 with 0.687 (see Tab. 2). The confidence interval (95% interval with one sample Wilcoxon Test) is [2.5:5]. The p-value is 0.01991 (Wilcoxon signed rank test). The precision of step 6 is the lowest of our method. The description of the safety goal was challenging for our participants and we have to provide further support for this effort in the future.

6.2. Precision Teams vs. Individuals: H_0^{PRET}

We illustrate the comparison between the results of the individual application of Steps 4 and 5 of our method and the application between groups in Tab. 3.

Table 3: Individuals vs. Teams

Step	**Individual**	**Group**
Step 4	0.843	0.937
Step 5	0.843	0.875

Step 4. With regard to our hypothesis, we can state that for Step 4 the value for groups is with 0.937 significantly higher than the value for the individual assessment with 0.843 (see Tab. 3). Hence, we can reject the hypothesis for step 4. We observed that the groups discussed the hazards intensively, but after a short time (less than 10 minutes) reached a unanimous agreement on what the most relevant hazard is and of what driving situations and faults it is comprised of.

Step 5. With regard to our hypothesis, we can state that for Step 5 the value for groups (0.875) is not significantly higher than for individual assessment (0.843; see Tab. 3). We observed that among the group discussions only minor improvements and in particular severity and exposure variables were discussed. We saw that the participants had problems with the examples in the standard text individually and could resolve these issues in the group discussions fast.

6.3. Productivity: H_0^{PRO}

We describe the results of our study with regard to productivity in Tab. 4. We present the missing (not instantiated) fields of the 8 participants in the first lines of the tables. Note that we did not check the correct instantiations, but we

included certain consistency checks such as if a driving situation is listed as not relevant another field should state in which situation it is included. MAX shows the maximal number of fields that should be instantiated. We describe the mean of the missing fields and the means of the instantiated fields, respectively. Finally, we show the productivity in percent. We report on the productivity results per step in the following.

Table 4: Evaluation of our Method: Productivity

Participant	Step2	Step3	Step4	Step5	Step6
1	4	0	0	0	0
2	0	2	0	0	0
3	0	0	0	0	0
4	4	1	1	0	0
5	4	0	0	0	0
6	4	0	0	0	0
7	5	1	1	0	2
8	2	0	0	0	5
MAX	20	18	8	8	6
Mean Missing	2.875	0.5	0.25	0	0.875
Mean Correct	17.125	17.5	7.75	8	5.125
Productivity	**0.856**	**0.972**	**0.968**	**1**	**0.854**

We have not achieved a 90% productivity in Steps 2 and 6. The remaining steps have only a small margin of not relevant uninstantiated fields. Hence, we focus our report on the steps 2 and 6.

Step 2. The productivity numbers confirm the problems with Step 2. Several participants did not fill relevant fields and only 2 of participants managed to instantiate all of the relevant fields. The not instantiated fields are mainly related to the relevance of driving situations. These problems are consistent with the ones reported in the precision section.

Step 6. The productivity drop in Step 6 is below the 90%, but the cause are only 2 participants. The remaining 6 has instantiated all relevant fields. In particular, the productivity drop is caused by participant 8 that did not instantiate 5 of 8 relevant fields. Thus, we do not believe this productivity drop demands a further investigation.

7. Debriefing

7.1. Exit Questionaire

We report in this section on the evaluation of our exit questionnaire. We asked the participants how they would evaluate themselves after having conducted the study. Figure 12 shows the results of that question. The blue bars (left) represent the assessment before the study and the orange bars (right) the

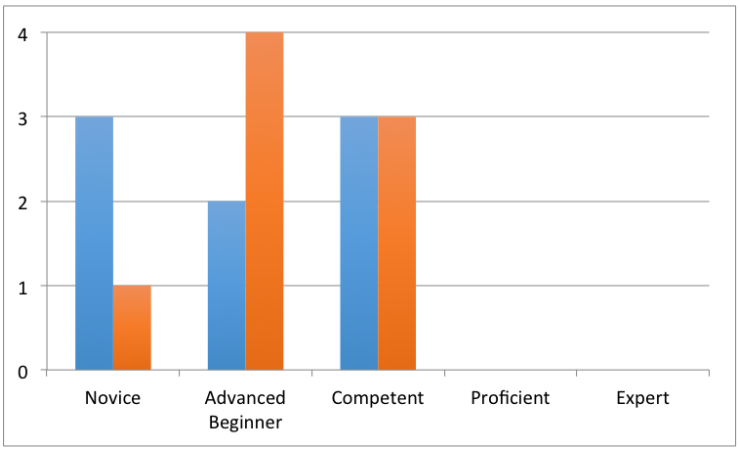

Figure 12: Experience in Safety Analysis Before/After

ones after the study. The evaluations shows that we had a shift from Novice to Advanced Beginner after the method.

Did you gain knowledge by applying our Method? The majority of our participants (6) simply stated that they gained knowledge by applying the method. One participant stated that the assessment is based on subjective criteria. Furthermore, one participant praised the structured method for reducing the effort for hazard and risk analysis. Additionally, the method prevents to overlook important cases.

Did you gain more knowledge by applying our Method individually or in a group? 7 participants rated the group discussions as an essential gain of knowledge. One participants mentioned the reason for this is the comparison of different opinions. 1 participant stated to have claimed more knowledge while working on the topic alone. However, even this participant admitted that the group discussions helped to gain confidence in risk values.

Are you satisfied with the result (ASIL and Safety Goal)? Would you sign with your name for the result, which may be used in a court of law? 7 participants answered with 'yes'. One of these participants added that the result equal his personal assessment. 1 participant stated that he requires more practice to make a decision.

7.2. Exit Interview

We asked all the participants to state positive and negative feedback on our method. We report first on the positive comments. The group discussions were a great experience. The initial work alone helped to understand the scenario and to prepare for the group discussion. The method matches with ones own intuition and the stepwise method is useful and comprehensible. Other methods that the participants had tried before did consider only situations and are missing the structured fault elicitation.

The comments for further improvement for Steps 2 and 3. Many participants stated that the decision criteria for building the hierarchies needs to be made more precise. Moreover, misunderstandings of the diagram in Step 1 are fatal, because they effect all subsequent steps. In addition, Step 5 was criticised. The examples in the standard for severity, exposure and controllability are insufficient. The participants asked for further examples to make better assessments. In addition, each step should come with an explanation stating why this information is needed for the next steps. Moreover, participants stated that it is hard to apply this method alone and that group discussions are essential to confirm assessments.

8. Threats to Validity

We disucss the threats to validity using the four validity classes proposed by Wohlin et al. [29].

Construct validity The measures made for the study in terms of precision and productivity might be not representative for measuring the applicability of the method. A possibility exist that measures could be identified that are better than ours. However, we investigated further analysis works in the field [27, 24] and did not identify any more precise methods than the ones used in this publications. In addition, we believe that the measure of the correctly instantiated fields (precision) is an intuitive best practice for template based approaches. Additionally, the amount of fields that were overall instantiated (productivity) shows if participants could possibly not fulfil the task of instantiating the field at all. This measure provides insides into significant issues with the approach. In addition, the final results achieved in groups instead of individual assessments may be influence by not only the group discussion, but the extra time the participants had to think about the hazards. Moreover, we mentioned to the students that we have a structured method and our interest in how well it can be applied. This may have influenced the participants. In addition, we limited the number of Hazop Guide words from 8 to 4, due to time constraints of the study. The 4 keywords were selected by the authors based on their expertise in the safety domain.

Conclusion validity The study was conducted on a paper in which each step of the method was outlined and some examples provided. The examples were perceived and applied well to a different context by some participants, while others stated that they would like more and different examples. We decided not to put too many examples in order to avoid too long texts. Moreover, the study was done on paper and digitalised into an excel file afterwards. We had two independent persons check that this digitalisation went without problems. In addition, we had only 8 participants in the study. However, all of these students are enrolled in a highly specific study path for automotive software engineering and had years of experience in working in the automotive domain. It is a particular challenge to gather a large sample group of this type, which is highly representative of the target audience.

Internal validity We had a single group of participants conducting the study. The reason is the focus on individual and group application of our method.

Moreover, a control group could have been given the ISO 26262 standard and applied it directly to the scenario. We decided against this approach, because the relevant parts of the standard span over many pages and the participants would have had to spend a significant time to read and understand the standard. We deemed this unreasonable in the time frame of the study. Moreover, we have a diffusion of treatment threat to validity. The study is based on a publication of our method and example [2] and the students could have read the paper before participating in the study. We provided the initial step of the method the item descriptions to the students. Hence, we cannot report on possible problems on creating an item description. In addition, we asked the students at the beginning of the study if they are familiar with that work and they declined. However, there is the possibility that some students may have answered less than honest. In addition, we used just one method in our test and did not compare the application of the scenario to multiple methods. The reason for this is that our aim was to figure out the general applicability of our method with junior safety engineerings and create a description of its application and identify steps that require improvement. Moreover, in industry engineers do have explicit team discussions for the hazard and risk assessment, but they have peer reviews in which experts from other departments or other companies redo the hazard analysis and discuss about one part to reach a common assessment. Finally, the researchers conducting the study are the same as the inventors of the method. This may have a positive effect on the qualitative evaluation.

External validity A threat to the external validity is that we conducted the study with students instead of practitioners from industry. However, the sample was taken in a specific master program for automotive software engineering and all of the students have had years of experience working in the automotive domain. Moreover, the students received many courses from practitioners from the industry that otherwise teach practitioners. Therefore, we are reasonably certain the results of this study provide reliable conclusions.

9. Conclusions and Future Work

We previously contributed a structured method for hazard analysis and risk assessment compliant to the ISO 26262 standard [2]. In this work, we contribute an empirical evaluation of our method with 8 participants from a master course in automotive software engineering at the Technical University of Munich (TUM). We provided the participants with a template, which fields they had to instantiate. All of the participants worked already in industry. We measured the successful application of our method in precision (true positives / number of fields). Furthermore, our participants applied the method initially alone and discussed the results afterwards in groups.

Our study revealed the following results:

- the participants could apply our method with expected results and in particular the hazard analysis and ASIL assessment worked reliably well (precision above 80%)

- our hierarchies for driving situations and faults worked to some extend but were challenging for the participants (precision above 70%)
- the greatest problem for the participants was the safety goal elicitation (precision above 60%)
- the group discussions improved the results of the hazard analysis (precision above 90%)

In the future, we will repeat the study with further students and practitioners. Moreover, we want to test the UML version of the method against the template-based version. We are interested in the effect UML models have on the comprehension of the participants when applying hazard analysis and risk assessment in the automotive domain.

References

[1] International Organization for Standardization (ISO), Road Vehicles – Functional Safety, ISO 26262 (2011).

[2] K. Beckers, T. Frese, D. Hatebur, M. Heisel, A Structured and Model-Based Hazard Analysis and Risk Assessment Method for Automotive Systems, in: Proceedings of the 24th IEEE International Symposium on Software Reliability Engineering, IEEE Computer Society, 2013, pp. 238–247. URL `http://www.ieee.org/`

[3] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), Functional safety of electrical/electronic/programmable electronic safety-relevant systems, IEC 61508 (2000).

[4] D. Tang, X. Qian, Product lifecycle management for automotive development focusing on supplier integration, Computers in Industry 59 (23) (2008) 288 – 295, product Lifecycle Modelling, Analysis and Management.

[5] S. Baumgart, Investigations on hazard analysis techniques for safety critical product lines, in: IRSCE12, ACM, New York, NY, USA, 2012.

[6] Safety Management System and Safety Culture Working Group (SMS WG), Guidance on hazard identification, Tech. rep., SMS WG (2009).

[7] P. H. Jesty, K. M. Hobley, R. Evans, I. Kendal, Safety analysis of vehicle-based systems, in: Proceedings of the 8th Safety-critical Systems Symposium, LNCS 1943, Springer, 2000, pp. 90–110.

[8] H. Mehrpouyan, Model-based hazard analysis of undesirable environmental and components interaction, Master's thesis, Linköpings Universitet (2011).

[9] H. Giese, M. Tichy, D. Schilling, Compositional Hazard Analysis of UML Component and Deployment Models, in: M. Heisel, P. Liggesmeyer, S. Wittmann (Eds.), SAFECOMP, LNCS 3219, Springer, 2004, pp. 166–179.

[10] Y. Papadopoulos, C. Grante, Evolving car designs using model-based automated safety analysis and optimisation techniques, Journal of Systems and Software – Special issue: Computer software & applications 76 (1) (2005) 77 – 89.

[11] A. Mouaffo, D. Taibi, K. Jamboti, Controlled experiments comparing fault-tree-based safety analysis techniques, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, EASE '14, ACM, 2014, pp. 46:1–46:10.

[12] L. Galvao Martins, T. De Oliveira, A case study using a protocol to derive safety functional requirements from fault tree analysis, in: Requirements Engineering Conference (RE), 2014 IEEE 22nd International, 2014, pp. 412–419.

[13] J. Jung, K. Hoefig, D. Domis, A. Jedlitschka, M. Hiller, Experimental comparison of two safety analysis methods and its replication, in: Empirical Software Engineering and Measurement, 2013 ACM / IEEE International Symposium on, 2013, pp. 223–232.

[14] T. Stålhane, G. Sindre, A comparison of two approaches to safety analysis based on use cases, in: Conceptual Modeling - ER 2007, Vol. 4801 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007, pp. 423–437.

[15] V. V. ans Michael G. Morris, G. B. Davis, F. D. Davis, User acceptance of information technology: Toward a unified view, MIS Quarterly 27 (3) (2003) 425–478.

[16] T. Stålhane, G. Sindre, Safety hazard identification by misuse cases: Experimental comparison of text and diagrams, in: Model Driven Engineering Languages and Systems, Vol. 5301 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2008, pp. 721–735.

[17] T. Stålhane, G. Sindre, L. du Bousquet, Comparing safety analysis based on sequence diagrams and textual use cases, in: Advanced Information Systems Engineering, Vol. 6051 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2010, pp. 165–179.

[18] R. Scandariato, J. Walden, W. Joosen, Static analysis versus penetration testing: A controlled experiment, in: Software Reliability Engineering (ISSRE), 2013 IEEE 24th International Symposium on, 2013, pp. 451–460.

[19] K. Yskout, R. Scandariato, W. Joosen, Do security patterns really help designers?, in: Proceedings of the 37th International Conference on Software Engineering - Volume 1, ICSE '15, IEEE Press, 2015, pp. 292–302.

[20] M. de Gramatica, K. Labunets, F. Massacci, F. Paci, A. Tedeschi, The role of catalogues of threats and security controls in security risk assessment: An empirical study with atm professionals, in: Requirements Engineering: Foundation for Software Quality, Vol. 9013 of Lecture Notes in Computer Science, Springer International Publishing, 2015, pp. 98–114.

[21] R. Slavin, J.-M. Lehker, J. Niu, T. Breaux, Managing security requirements patterns using feature diagram hierarchies, in: Requirements Engineering Conference (RE), 2014 IEEE 22nd International, 2014, pp. 193–202.

[22] D. Chen, R. Johansson, H. Lönn, Y. Papadopoulos, A. Sandberg, F. Törner, M. Törngren, Modelling support for design of safety-critical automotive embedded systems (2008) 72–85doi:10.1007/978-3-540-87698-4_9.
URL http://dx.doi.org/10.1007/978-3-540-87698-4_9

[23] V. R. Basili, G. Caldiera, H. D. Rombach, The goal question metric approach, in: Encyclopedia of Software Engineering, Wiley, 1994.

[24] M. Eraut, Developing Professional Knowledge and Competence, Falmer Press., 1994.

[25] Road vehicles – Controller area network (CAN), ISO 11898 (2003).

[26] IEC, Hazard and Operability Studies (HAZOP studies), ISO/IEC 62882, International Electrotechnical Commission (IEC) (2005).

[27] R. Scandariato, K. Wuyts, W. Joosen, A descriptive study of microsoft's threat modeling technique, Requir. Eng. 20 (2) (2015) 163–180.

[28] D. Holling, Course on Automotive Software Engineering, Tech. rep., Technical University of Munich (TUM) (2015).
URL http://www.in.tum.de/fuer-studieninteressierte/master-studiengaenge/automotive-software-engineering.html

[29] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in Software Engineering, Springer, 2012.