

Practical Challenges of Requirements Prioritization Based on Risk Estimation: Result of Two Student Experiments

Technical Report SWEHD-TR-2008-03



Authors:
Andrea Herrmann
Barbara Paech

Version 1.3
16th July 2008

**A publication of the
Software Engineering Group
University Heidelberg, Germany**

The „Software Systems Engineering“
Group is part of the Institute of
Mathematics and Computer Sciences of
the Ruprecht-Karls-Universität
Heidelberg. It is managed by

Prof. Dr. Barbara Paech
Institut für Informatik
Neuenheimer Feld 348
69120 Heidelberg
Germany

paech@informatik.uni-heidelberg.de
<http://www-swe.informatik.uni-heidelberg.de/>

Document History

Version	Date of Change	Author	Change
Version 1	November 2007	Andrea Herrmann	Initial version
Version 1.1	January 2008	Andrea Herrmann & Doris Keidel-Müller	Revision of spelling and of style
Version 1.2	July 2008	Andrea Herrmann	Inclusion of tests of statistical significance of differences

Abstract:

Requirements prioritization and risk estimation are known to be difficult. However, so far no quantitative empirical investigation has investigated how risk-based requirements prioritization can be improved. We performed two quantitative experiments about this principle. Our aim was to explore the practical challenges and needs of risk estimations in general and of our method MOQARE specifically. In the first experiment, ten students did individual estimations. In the second one, twenty-four students estimated risk in seven moderated groups. The students prioritized the same requirements with different methods (risk estimation and ranking). During the first experiment, we identified factors which influence the quality of the prioritization. In the second experiment, the results of the risk estimation could be improved by discussing risk estimations in a group of experts, gathering statistics about probabilities and damages caused by risks, defining requirements, risks and prioritization criteria more tangibly. This first quantitative study about risk-based requirements prioritization helps to understand the practical challenges of this task and thus can serve as a basis for further research about this topic.

1.	Introduction	7
2.	Basics of Requirements Prioritization and Risk Estimation	8
2.1	Requirements Prioritization Based on Risk	8
2.2	Empirical Studies on Requirements Prioritization	11
3.	The Research Questions, and the Requirements Prioritization Methods Used in the Experiments.....	12
4.	Experiment No. 1	14
4.1	Experiment No. 1: Preparation.....	14
4.2	Experiment No. 1: Execution	16
5.	Experiment No. 2	16
5.1	Experiment No. 2: Preparation.....	17
5.2	Experiment No. 2: Execution	19
5.3	Influencing Factors in Experiment 1 and 2	20
6.	Results and Lessons Learned	21
6.1	Influence of Statistics	21
6.2	Improvements Between Experiment 1 and 2	22
6.3	Influence of Moderated Group Discussion	23
6.4	Influence of Transparency.....	25
6.5	Prioritization Criteria and Metrics.....	25
6.6	Order Effects and Learning Effects.....	26
6.7	Comparison of the Methods	27
6.8	Further Lessons Learned	29
6.9	Discussion of Validity	32
7.	Conclusion and Future Work	34
8.	References	36
	Annex A: Experiment Material	39
A.1	Experiment 1	39
a)	Presentation slides	39
b)	Material: Handouts and Questionnaires.....	46
c)	Spreadsheet tables	80
A.2	Experiment 2	87
a)	Presentation slides.....	87
b)	Material: Questionnaires	95
c)	Spreadsheet tables	103
	Annex B: Data and Data Analysis.....	107
A.1	Time Consumption	107
A.2	Priorities	108
A.a	Standard Deviation of Priorities	108
A.b	Ease of Use.....	109
A.c	Participants Expect their Estimations to Be Realistic	109
A.d	Accuracy of the Results.....	109
A.e	Uncertainty of the Estimated Values and of Countermeasure Benefits	110
A.f	Frequency of Naming a Misuse Case or Countermeasure as Especially Uncertain	110
A.g	Participants Feel Certain	111
A.3	Influence of Statistics	112

1. Introduction

Requirements prioritization is known to be difficult to perform: “Requirements decisions are hard because of the uncertainty and incompleteness of the information available.” (Ngo-The, 2005) There are even more factors which amplify this difficulty, e.g. differing perspectives of the stakeholders and dependencies among requirements. One way of identifying priorities of requirements is to assess the risks involved in the case a requirement is not realized (Berander, 2004),(Park, 1999). This principle is being used especially in the context of security (like in (Arora, 2004)), but also makes sense with other non-functional requirements (NFR) (Feather, 2006) when prioritizing countermeasures. Countermeasures are a special type of requirements which are defined in order to reduce risk and by doing so improve software quality. We have applied this principle in MOQARE (Misuse-oriented Requirements Engineering) (Herrmann, 2005),(Herrmann, 2006a),(Herrmann, 2007), which is a systematic method for the elicitation of countermeasures. In MOQARE, countermeasures are directly derived from misuse cases and indirectly from business goals. Based on the quantification of the misuse case risk, the risk reduction achieved by the countermeasure can also be quantified. This information can be transformed into priorities.

There are no publications of quantitative empirical studies on risk-based requirements prioritization. Therefore, we performed two student experiments to learn about the practical needs of this requirements prioritization principle, like preparation, material, knowledge and time consumption. We wanted to understand the factors which influence the quality of the outcome. The insights gained during the first experiment helped to improve the outcome of the second experiment.

Before performing these experiments, our prioritization approach had been tested in examples and in one case study. However, the resulting impressions of the quality of the method were subjective and depended on the estimator and the context. Therefore, we decided to make a systematic empirical investigation with a considerable number of participants who all perform the same task, who have a comparable educational background, level of knowledge of the method and of the case system used in the experiment. The participants had to be open for experiments and for using different methods to solve the same problem just for the sake of testing them. The experiment was performed in a lecture. So, several parameters could be controlled, e.g. which material is used, and whether the participants perform the tasks in the defined order. We did not expect to find such participants and conditions in a real software project. Therefore, conducting a student experiment seemed ideal for our purpose of performing a pilot study. A case study in a real project or further experiments will be the next steps.

The remainder of this technical report is structured as follows: In section 2, we summarize some basics of requirements prioritization and risk estimation about prioritization, including published empirical investigations. Section 3 describes the requirements prioritization methods tested and the research questions, the variables which were observed in the experiments. Section 4 describes the preparation and execution of experiment 1. Section 5 describes experiment 2 in the same form. In section 6, the results of both experiments are being discussed and compared and lessons learned derived. In section 7, we discuss our conclusions and future work. Annex A displays experiment material of both experiments, and Annex B summarizes the data and data analysis for both experiments.

2. Basics of Requirements Prioritization and Risk Estimation

Before describing our own requirements prioritization method and the experiments, we cite some approaches of requirements prioritization, especially the ones that use risk estimations. Then we present former empirical work about requirements prioritization.

2.1 Requirements Prioritization Based on Risk

Typical criteria for the prioritization of requirements are their benefit (e.g. business value) for the stakeholder, the dissatisfaction if not implemented, their urgency, volatility, risk, their implementation cost or system impact. They can be estimated in cardinal values (=absolute values) or ordinal values (=relative values, also called ratio scale). There are several methods for determining ordinal rankings, often based on pair-wise comparison, mainly varying in the way the pairs are being combined. Karlsson et al. (Karlsson, 1998) describe and compare six such methods: analytic hierarchy process (AHP), hierarchy AHP, minimal spanning tree matrix, bubble sort, binary search tree, priority groups. Another method is the 100\$ method (also called cumulative voting) (Leffingwell, 2000). Only few of these methods can be used for the determination of cardinal values. Risk estimation is such an approach. An advantage of priorities on a cardinal scale we see in their scalability and extensibility. New requirements can easily be inserted in an existing list of prioritized requirements, without the need of comparing them to the whole list of the other requirements.

Misuse cases describe the course of risk events (e.g. attacks, user errors, accidents) which happen with a certain probability and have a usually negative impact. Unlike the traditional use for eliciting security requirements (Sindre, 2000), (Sindre, 2001), we apply misuse cases to all types of NFR and therefore these misuse cases describe as well unwanted scenarios, where the misuser is not a malicious attacker, but might be a user who by mistake impairs data integrity or where a developer's negligence threatens the maintainability of a software. Misuse cases are used to identify countermeasures, i.e. requirements which, if satisfied, prevent, mitigate or detect misuse cases and by this support the satisfaction of security requirements and of other NFR. Countermeasures can be requirements on the IT system, on its design, its development process, operation environment or personnel.

Risk events have an effect on the benefit and cost of the system. In the security community, the risk of misuse cases is quantified by the product of *probability* and *caused damage* (see for instance in (Kontio, 1996), (ISO, 2002), (Xie, 2004)). This risk is influenced by which countermeasures are realized or not, but also depends on environmental factors.

Regularly, risk is proposed to not only quantify the importance of misuse cases, but also the benefit of a system or of a single requirement. In (Park, 1999), WinWin is described to "assign each item [=requirement] a difficulty and importance (or a probability and loss)". Berander (Berander, 2004) explicitly uses risk estimations as a prioritization criterion. Mayer, Rifaut and Dubois (Mayer, 2005) propose to integrate requirements engineering and risk analysis "for focusing [...] on the most critical parts of the IS." They use the quantitative risk assessment, business criticality, budget and the countermeasure cost as a basis for the requirements elicitation and prioritization.

The Failure Mode and Effects Analysis (FMEA) (Stamatis, 2003) also prioritizes failures according to risk, which in FMEA is defined as the product of the importance of the failure effect, the probability of occurrence of the failure cause and the inability of controls to detect the failure effect or failure cause. Each of these three aspects is rated by a number between

one and ten, which results in a risk between 1 and 1000. (Other approaches estimate probabilities in percent and damages in \$.)

Feather et al. (Feather, 2006) measure the benefit of a countermeasure (there called “mitigation”) by the difference between the risk without any countermeasures being implemented (worst case) and the risk with the chosen countermeasure. Arora et al. (Arora, 2004) also explicitly set the benefit of a countermeasure equal to “the reduced expected loss due to security failure incidents (i.e. reduction in risk)”. To determine this risk reduction, they define two types of risk: the baseline risk and the residual risk. The baseline risk is calculated from total incident risks, if no countermeasures were in place. The residual risk is the expected value of damages, if only one countermeasure was installed. Then, the benefit of this countermeasure equals baseline risk minus residual risk. These authors do not refer to a real case study, only to examples, where supposed values are used.

Although the estimations of risk and risk reduction have often been proposed, we found only two in-depth experience reports of authors who actually applied risk estimations. In the SQUARE project (Xie, 2004), Xie et al. applied the quantitative principles of Arora et al. (Arora, 2004) in small companies, but met several practical challenges. They remark that this approach has a practical limitation: It requires high volumes of incident data, ideally from the same company. While big companies generate their own historical security statistics, small companies must rely on public statistics. They report: “detailed attack data are simply not available to be used as references”. As public statistics are only available on a high level of granularity, Xie et al. (Xie, 2004) subsume misuse cases in categories of threat, such as denial of service, system penetration, or sabotage of data. Similarly, countermeasures are summarized in categories. Xie et al. define the baseline risk like Arora et al. (Arora, 2004), but the residual risk as “incident risk to the organization if security solutions are properly installed, utilized, and monitored”. They initially used estimated cost figures from nationally surveyed losses for each category of threats. Later on, they worked with a company and their estimations for their environment. They found that lower ends of nationally surveyed losses may be used as estimations for tangible losses (productivity loss, fixing cost, etc.), but cannot sufficiently account for intangible losses (loss of reputation, loss of confidential data, etc.), since these values are highly company and project specific.

The second group of experience reports stems from the NASA. Feather et al. (Feather, 2006) as well performed a high number of risk estimations, but they do not discuss practical challenges. They emphasize the importance of visualization. Their positive experiences from many case studies show that risk estimations make sense and can be applied successfully.

While there is only little practical experience with risk estimation in the field of requirements prioritization, it is different in the decision theory community. Many biases are known (Raiffa, 2002) which lead to bad estimations of probabilities, frequencies and values, or as Tversky and Kahneman (Tversky, 1974) put it: “intuitive predictions and judgement under uncertainty do not follow the laws of probability or the principle of statistics. Instead, people appear to rely on a limited number of heuristics and evaluate the likelihood of an uncertain event by the degree to which it is representative of the data generating process, or by the degree to which its instances or causes come readily to mind.”

Another handicap of requirements prioritization is, that existing methods for requirements prioritization do not consider dependencies among the benefits of requirements at all or only superficially. Such methods are the analytic hierarchy process (AHP) (Saaty, 1980), (Karlsson, 1998), numeral assignment (Karlsson, 1996) or cumulative voting (CV), also called “\$100 test” (Leffingwell, 2000), (Berander, 2006). According to (Herrmann, 2006), all methods which attribute a fixed priority value to a requirement can be said to neglect dependencies.

In reality, however, such dependencies are critical (Ryan, 1997). In one of our publications (Herrmann, 2006), we discuss how such dependencies complicate estimations. For instance,

countermeasures can replace each other partly, when they mitigate the same misuse case. The benefit of implementing both of two such countermeasures is not twice as high as the benefit of only one, but less. Then again, two or more countermeasures may also need each other for being effective against a misuse case. In this case, the implementation of one of these countermeasures alone does not add much benefit, only the implementation of all them. Due to such dependencies, the benefit of a requirement cannot be described by one fixed value only (as is usually done) and also, several benefits cannot be added up. Instead, the benefit of a group of requirements must be estimated as a whole. In some prioritization methods, it is common to bundle those requirements, which depend on each other most, in relatively independent bundles. These bundles have the name feature (Regnell, 2001), [Wi99], feature group (Regnell, 2001), super-requirement (Davis, 2003), class of requirements (Ruhe, 2003), bundle of requirements (Papadacci, 2004), category (Xie, 2004), User Story (Beck, 2000), super attribute (Stylianou, 1997) or Minimum Marketable Feature (Denne, 2003). Bundles are applied as an efficient way of reducing the complexity, time need and effort of prioritization. However, bundling considers only the strongest dependencies.

In order to take into account dependencies, it is also important to relate all risk and benefit estimations to the same reference system (Herrmann, 2006). A reference system is an idea of a set of requirements which are imagined to be implemented. It is important that the reference system is clearly defined, easy to imagine for the estimators and near to the system that is finally to be implemented. If perfect quality is the goal or the benchmark, the perfect system is the reference, i.e. the system in which all countermeasures are implemented (Xie, 2004). The reference system can also be the ensemble of all mandatory requirements (Ruhe, 2003), the former system version, a competitor's product or all FR without any countermeasures (Arora, 2004), (Feather, 2006). The reference system can take into account more complex dependencies than bundling. The "reference risk" denotes the risk in this reference system (like the residual risk in (Xie, 2004)). In order to determine the risk reduction effected by the implementation of each countermeasure relative to the reference system, we estimate a "varied risk" in a system identical to the reference system, except for one countermeasure only.

The risk reduction respectively benefit achieved by a countermeasure in relation to a misuse case equals the difference between the two misuse case risks: the reference risk and the varied risk (Arora, 2004). The estimation is complicated by n-m-relationships between misuse cases and countermeasures. If one countermeasure counteracts several misuse cases, then the benefit of the countermeasure can be the sum of the risk reductions produced for each of these misuse cases, but not necessarily, if dependencies exist among the misuse cases. To account for such dependencies, a misuse case group can be estimated together. If vice versa for one misuse case several countermeasures are defined, the estimation of the risk reduction is done for each misuse case – countermeasure pair individually, each time in a system which differs from the reference system by variation of this single countermeasure alone. If the countermeasures depend on each other strongly (either by being able to replace each other partly or totally, or by being only effective if implemented together), the countermeasures are bundled and treated as one countermeasure. Bundling can avoid effects like this: When several countermeasures can replace each other, then the benefit achieved by each of them relative to the perfect system is low (and consequently its priority), although the whole bundle may be highly beneficial. Consequently, these countermeasures would be prioritized too low.

2.2 Empirical Studies on Requirements Prioritization

So far, there have been no quantitative empirical studies about risk-based requirements prioritization. Therefore, in this section, we describe four empirical studies on other requirements prioritization methods. We will build the design of our own experiments on these studies. We could not find other experiments which were as thoroughly designed, executed and analysed. Other publications on empirical experience in requirements prioritization instead refer to qualitative industry case studies and treat mostly organizational aspects of requirements prioritization, like information flow, stakeholders involved or negotiation of different opinions. However, we are highly interested in systematic and controlled experiments, like we plan to perform.

Karlsson (Karlsson, 1996) performed an empirical comparison of the *pair-wise comparison technique* and a *numeral assignment technique* with five participants, applying them on 14 requirements. Criteria for the comparison were time consumption, number of comparisons to execute, standard deviation of the priorities for the same requirement, perceived trustworthiness of the method. They found that relative prioritization by pair-wise comparison of requirements and judging which is more important relative to the other, tends to be more accurate and informative than attributing absolute numbers to the requirements. Relative values were also found to be easier to estimate than absolute values.

Karlsson, Wohlin and Regnell (Karlsson, 1998) compared six prioritization methods in a self-experiment. Each of the three authors prioritized the same 13 quality requirements. Their criteria for the comparison were: number of decisions, time consumption total and per decision, ease of use, subjective reliability of results, fault tolerance. They concluded that *AHP (Analytical Hierarchy Process)* is best, because it produced the most trustworthy results, is fault tolerant, includes a consistency check, and the distance between requirements becomes tangible. Its main problem is scalability: the time consumption grows with the square of the number of requirements.

Karlsson et al. (Karlsson, 2004) describe an experiment aimed at comparing the *Planning Game PG* with *Pair-Wise Comparison*. They measured the average time consumption and assessed the ease of use by asking: “Which technique did you find easiest to use?”. The accuracy was measured in a post-test a few weeks after the experiment. The subjects were asked which of the two resulting priority orders reflects their opinion best. The experiment was performed with 15 Ph.D. students and one professor as subjects. They prioritized features of mobile phones with respect to both prize and value. The results indicate that PG is less time-consuming and a majority of the subjects found it easier to use. Most subjects also found the results from PG more accurate, i.e. they said that they reflect their views more accurately, which was unexpected. To find out whether order effects occurred, the two techniques were performed in varied order, the aspects prize and value were treated in different order, and also the requirements were presented in different order. However, no statistically significant order effect was observed.

Karlsson et al. (Karlsson, 2007) performed a further experiment for comparing *tool supported pair-wise comparison* with the *PG*. They observed the same variables and used the same requirements as above. Half of the subjects were asked to prioritize 8 requirements, while the other half prioritized 16 requirements. There were no statistically significant differences between the results for 8 or 16 requirements, for instance no fatigue effect. The order in which the techniques were used affected the mean consistency ratio, but not to a statistically significant degree. When being tool-supported, Pair-Wise Comparison required less time than PG. PG seemed to be less easy to use and its results to be slightly more accurate, but the differences were not statistically significant.

From these experiments, we conclude for our own experiment, that

- a number of 8 to 16 requirements can allow conclusions on the properties of prioritization methods.
- interesting variables for comparing risk estimation methods are: the number of estimations to be done, the time consumption, the standard deviation of priorities for the same requirement, the subjectively perceived ease of use, and the accuracy, i.e. whether the participants think that the resulting priorities reflect their view; fault tolerance does not apply to our method as no redundant estimations are foreseen.
- we rather make the participants estimate relative than absolute values wherever possible.
- we have to discuss or vary the order in which different methods are performed.

3. The Research Questions, and the Requirements Prioritization Methods Used in the Experiments

As there are very few publications of empirical studies about estimations of risk and risk reduction, we performed two student experiments to empirically investigate risk estimations and the practical needs of this principle: How much time does it take, what knowledge is needed (e.g.: How are estimations influenced by statistics provided?), how much method training and what material do the estimators need, what is the expected uncertainty of the resulting values, how do group decisions influence the process and results, what is the influence of transparency? (Transparency in the context of our methods means: While estimating misuse case probabilities and damages, the estimator can see their effect on the resulting requirements benefits.) We also investigated the advantages and disadvantages of risk estimation compared to a simpler prioritization method.

These research questions were investigated by evaluating qualitative comments of the participants and observations of the moderators as well as by comparing the quality of the risk estimations under different conditions (e.g. when using different methods, with or without risk statistics provided, with or without transparency, with or without group discussions).

To investigate the advantages and disadvantages of requirements prioritization on the basis of risk estimations, we executed the same task with three different methods: with two risk estimation methods and with a traditional prioritization method.

To have a reference method to which to compare risk estimation, we chose a very simple prioritization method: ranking (described in more detail below). It is called “Method 1” in this work. We did not choose AHP as reference, although according to the experiments mentioned in section 2.2, it seems to be the best prioritization method available. However, our purpose was to investigate the benefits practitioners would have by performing risk estimation. Assuming that they usually do not choose between one sophisticated method and the other, but rather between a sophisticated method (like risk estimation) and a simpler one, we wanted to simulate this comparison.

We wanted to investigate two alternative ways of estimating the benefit of countermeasures in terms of the risk reduction which they produce: One can derive the risk reduction gained by a countermeasure directly from the corresponding misuse case’s risks as described in section 2.1. We call this “Method 2” in what follows.

But one can also profit of MOQARE’s characteristic that countermeasures can be traced back to business goals. Therefore, “Method 3” can derive the damage caused by a misuse case from the business goal’s benefit. One of the fundamental principles of MOQARE is to not only specify wanted behaviour but also unwanted effects like misuse cases. The wanted elements are quantified by their benefit, the unwanted by their risk. In MOQARE, the

following concepts are linked to each other in this order: business goal – business damage – quality deficiency (of the IT system) – quality goal – misuse case – countermeasure. In experiment 1 (where method 3 was applied), our reference system for estimations is the system where all countermeasures are implemented. For estimating the benefit of the business goal relative to this reference system, one asks what loss one would experience if the business goal (e.g. “to make profit”) was not achieved at all. In the second step, one asks how each business damage can harm the business goal and to which degree. We estimate this degree in percentages, i.e. in relative values, which - as was said above - is easier to do than estimating absolute values. Then, one estimates with which probability each quality deficiency leads to the business damage (e.g. when there are usability deficiencies, with which probability does this lead to a lower market share?). If the quality goal is not achieved, how much of the quality deficiency is caused (for instance if the user interface is badly learnable, to which degree the quality deficiency “usability deficiencies of the system” is caused)? When by the above procedure the benefit of a quality goal respectively the loss by its non-satisfaction is known, all damages caused by a misuse case can be estimated relative to the quality goal’s benefit. The benefit of the countermeasure is calculated from the misuse case risks as in method 2.

So, there were three methods to test:

1. Do an **intuitive ranking** in two steps: First, attribute each requirement to one of the groups “high/ average/ low benefit”, and then perform a total ranking, attributing the number 1 to the most beneficial requirement and the highest number to the least beneficial one.
2. Estimate the **reference risks** ($p_{ref} \times d_{ref}$) and **varied risks** ($p_{var} \times d_{var}$) of the misuse cases and from these calculate the benefit of each countermeasure to be $|p_{var} \cdot d_{var} - p_{ref} \cdot d_{ref}|$. The countermeasure with the highest benefit gets the highest priority 1.
3. Estimate the **benefit of the business goals** first and then the risk of the business damages, etc., down to the risk of the misuse cases and the benefit of the countermeasures as described above.

Method 3, the more complex approach, on the one hand is expected to allow more realistic damage estimation than method 2, on the other hand, many more estimations are necessary, which demand time, and the consecutive multiplication of estimates – each flawed with an uncertainty - raises the total uncertainty of the result. To find out whether the higher effort is remunerated by a better quality of the resulting requirements priorities, we **compare the methods**.

In the sections 4 and 5, both experiments will be described in more detail, and at the end of section 5, the parameters varied in the experiments are summarized in Table 1.

The influence of these factors is analyzed. We expected effects with respect to the time need, the quality of the resulting estimations and priorities, as well as to the participants’ subjective perception of this process and its results. In the remainder of this section, we present the variables used to measure these effects.

The **time need** of the methods is measured by the average duration in minutes the experiment participants need for their execution. As in practice, time need means cost, this variable is relevant for practitioners.

We measure the **quality of the priorities** resulting from the methods with the following variables (all subjective, except for the first one):

- a. a low **standard deviation** of the priorities of each single requirement (calculated over all participants’ or all group results), averaged over all requirements
- b. the participants indicate in a questionnaire that the method was **easy to use**

- c. directly after having done the risk estimations, the participants expect that the resulting priorities will be reasonable, **realistic** and useful (in experiment 1 without knowing them yet; in experiment 2 they know the resulting priorities)
- d. **accuracy**, i.e. the priorities resulting from the risk estimations reflect the participant's opinion
- e. **certainty**, i.e. the interval of estimated values as expected by the participants directly after the estimations, in %
- f. a low **frequency** with which a countermeasure or misuse case is named by the participants when being asked where they believe that their estimations were especially uncertain
- g. directly after the estimations, the participants **feel certain** about their estimations

In method 3, the damage estimations are expected to be easier to do than in method 2. (The damage estimation is expected to make the main difference between these two methods, because the probability estimations are identical.)

Additionally to these quantitative results, we also explicitly asked the participants to comment on the methods and the influence of group discussions in free-text fields and we gathered observations made by the experiment moderators.

4. Experiment No. 1

During experiment 1, nine countermeasures were prioritized by ten participants in individual estimations with the methods 1, 2 and 3.

4.1 Experiment No. 1: Preparation

Sample population: Ten master students participating in our university course “Knowledge management and decisions in software engineering” in the winter term 2006/07 performed this experiment in a three hour session. They had been taught prioritization methods and how to prioritize countermeasures within MOQARE in the lecture before, three weeks ago.

Requirements: The experiment uses a case study discussed in the lecture and in preceding homework: an Internet flea market to sell used goods from individuals to individuals. During the lecture, the business goal, business damages, quality deficiencies and quality goals had been identified, during their homework, the students described functional requirements, misuse cases and countermeasures. A consolidated list of all these requirements was discussed in the subsequent lecture. The homework had been done by all participants. Five participants had been participating in the discussion of the case study as well as the homework results, and five missed one or both. This known case study was chosen in order to reduce misunderstandings about the software system and its environment. We expect that the discussions in the lecture and the homework had almost no influence on the experiment results as risks, benefits and priorities had not been discussed there. Nine of the countermeasures defined during the homework were chosen for the experiment. The criteria for this choice were: They tackle a tangible risk which is easy to imagine, and they belong to different quality attributes, i.e. security requirements as well as usability requirements and maintainability.

Each of the methods was tested on the same nine countermeasures of the case study. These nine countermeasures were:

- **R1:** clear and intuitive user interface design
- **R2:** user support via several media (phone, email)

- **R3:** similarity with a real life flea market
- **R4:** inspection of the specification documents
- **R5:** encrypted storage of the customer data
- **R6:** fast hardware and software
- **R7:** standard compliance during its implementation and of the user interface design
- **R8:** automated notification of service staff in case of a system breakdown
- **R9:** backup server

(Although some of these countermeasures sound like design elements, we treat them as requirements, because it is unclear whether they will be satisfied by the system; they might not.)

In experiment 1, we define the reference system to be the system in which all countermeasures are implemented. Other reference systems could have been used, but this one is easy to imagine and easier to handle than a reference system in which some requirements are implemented and some are not. In this specific reference system, the varied risk is the risk in a system where all countermeasures are implemented, except for one.

Material: The material was pre-tested in a self-experiment by the two authors and by a colleague. The pre-tests had led to simplifications of the questionnaire and to an improved presentation of the methods.

During the experiment, the students individually estimated risks on paper questionnaires. They were led by step-by-step instructions and templates. We reduced the number of estimations to be done to the necessary minimum. All calculations and the derivation of benefits and priorities from the risk estimations were done afterwards by us. This was supposed to reduce the time consumption of the experiment, and also to allow to test the accuracy under intransparency, i.e. how well the priorities resulting from risk estimations correspond to the participants' view when (s)he cannot predict which effect her/ his probability and damage estimation(s) have on the resulting countermeasure priorities. The participants worked with the following six documents:

- The introduction which describes the objective of the experiment and the case study, including information about competitors, the company, project execution and staff, and the flea market's functional requirements.
- Questionnaire Q1 supports method 1 by a table which allows to attribute a group and a priority to each countermeasure. It also asks how certain the participant feel about their judgements (variable g).
- Questionnaire Q2 supports method 2. It first states the expected revenues and costs of the system and informs that all estimations are to be done for a period of two years. The reference system is defined to be the perfect system where all countermeasures are supposed to be implemented. Method 2 is supported by two tables: one for the reference risk and the other for the varied risk, each containing a column for the probability and for the relative damage estimation. Neither misuse case risk nor countermeasure benefit are calculated here. The participants are asked what relative uncertainty (in % points) they expect for their probabilities and damages and which were misuse cases where they were especially uncertain (variables e and f).
- Questionnaire Q3 supports method 3. Each of the steps described in section 3 is supported by one separate table. Here, exclusively relative values are estimated for probabilities and damages. Estimated values are reused from questionnaire 2 as far as possible. Like in Q2, the participants are asked what relative uncertainty (in % points) they expect for their probabilities and damages and whether there were misuse cases where they were especially uncertain (variables e and f).
- Questionnaire Q4 directly after the experiment asks the participants to rate the methods in terms of ease of use (Q4a, variable a) and whether they expect reasonable, realistic and useful results (Q4b, variable c).

- Questionnaire Q5: One week after the experiment, during the post-test session, each participant receives a table with his/ her priorities resulting from each method. They are asked the results of which method reflect their opinion best (Q5a, variable d). Question Q5b asked whether the damage estimation was easier in method 2 (where absolute values are estimated) or 3 (relative values), or equal. Q5c asked how certain the participants feel concerning their estimations in each method. Question Q5d offered four statistics on the frequencies of security incidents and the sources of attack taken from the CSI/FBI Computer Crime and Security Survey (Richardson, 2003). On this basis, the reference risk probabilities of two security misuse cases were re-estimated (the participants could look up their former estimations if they wanted to). The participants were asked by Q5e whether and how the statistics facilitated the estimations. Question Q5f showed the resulting priorities of all participants, including averages and standard deviations and asked what might be causes of these deviations.

4.2 Experiment No. 1: Execution

The methods 1, 2 and 3 were performed in this order. We did not vary the order, because we expected that performing the more sophisticated method first and then the more primitive one would influence the results of the latter, i.e. that for the simple ranking in method 1, risks would be taken into account to a higher degree than if method 1 was applied first. In the first experiment, we wanted to avoid such an effect. (Such order effects were tested in the second experiment.)

Before the first questionnaire was distributed, there was an introductory presentation which explained the objective of the experiment and the time plan, as well as the case example. After questionnaire Q1 and before Q2, the principles of countermeasures prioritization by risk estimation and of the reference system were recapitulated.

The experiment was performed in a three hour session, with 10 participants. Concerning the risk estimations in Q2, a discussion arose about how to interpret percentages in the probability estimations. For instance, for some misuse cases the probability means “What ratio of the users...?”, for others: “How long during two years...?” Some information was found to be missing in the material, like: When a potential buyer does not buy an article, how high is the probability to find another buyer?

In two cases, in Q2, there evidently had been misunderstandings during the risk estimations: When estimating the damage for the varied risk, a value relative to the reference risk damage was to be given, not relative to the value of the business goal, as in the estimation of the reference risk before. This misunderstanding had the consequence that the risk with the countermeasure being implemented was higher than without. These two participants received their results in the form of excel spread sheets per email and corrected their varied risk damage estimations before the post-test session.

One week after the experiment session, a post-test was performed where questionnaire Q5 was answered by the participants individually and afterwards a concluding discussion of the experiment took place in the group.

5. Experiment No. 2

During experiment 2, seven countermeasures were prioritized in moderated group estimations with methods 1 and 2.

5.1 Experiment No. 2: Preparation

Sample population: 24 students took part in this experiment in the summer term 2007. Three groups had 4 members and four groups had 3 members. These groups were identical to the teams which had formed five weeks earlier and worked on a programming project together in the software engineering course. In these teams, almost all members were bachelor students of the “Software engineering I” course. Except for one 3 person team, all groups included a project manager, i.e. a master student (a bachelor student in one case) taking part in the course “Software engineering II: Requirements Engineering and Project Management”. One project manager did not take part in the experiment, because this student had before been a participant in experiment 1 and we therefore expected a strong influence of his risk estimation experience in the discussion. The moderators were advised to treat all team members equally.

The students had not been taught prioritization methods or how to prioritize requirements in MOQARE before, but received a twenty minutes introduction at the beginning of the experiment.

The experiment was performed in sessions of one and a half hour.

In this second experiment, some unwanted factors from the former experiment were eliminated. These factors are discussed in section 5.3. For instance the students estimated risks for a system which they know well because they had been using it in this course for at least four weeks and were enhancing its code. Consequently, their estimations were based on user knowledge and programmer knowledge of a real system. Furthermore, no detailed description of the system and its environment needed to be provided by us and to be understood by the participants. Estimates were to be discussed in groups, in order to reduce the probability of misunderstandings. Unlike in the first experiment, the consequences of estimations on the final priorities were made transparent. The measures with which probabilities and damages were estimated, were defined in a more tangible way. Probabilities were estimated in “times per month in average” instead of percentages. Damages were measured in “lost calendar hours” instead of %. In experiment 1, method 1 was executed without defining clear prioritization criteria. In experiment 2, the priority was defined to be the benefit with respect to the misuse case (without estimating the risk quantitatively). To observe the effects of these changes is one of the objectives of this experiment 2.

Requirements: The software we used as an example, was Sysiphus (Dutoit, 2002), (Dutoit, 2003), (Wolf, 2004), (Sysiphus, 2007), a software engineering tool which was developed at the Technical University of Munich and at the University of Heidelberg. Both universities use the tool Sysiphus to teach software engineering and to document the results of case studies. It serves two purposes: It supports the whole software engineering process in student projects and it serves as a realistic example software which students modify to learn programming. Five weeks before the experiment, as a homework, the students had derived misuse cases and countermeasures for the quality goals “conformance of user interface to user expectations”, “performance”, “availability”, “compliance to Java Code Conventions” and “clear code structure”. Some of the misuse cases and countermeasures were then used in the experiment. The students consequently knew MOQARE and had thought about misuse cases and countermeasures, but had not learned or thought about prioritization before.

The countermeasures prioritized in this experiment were:

- **R1:** usability tests with future users
- **R2:** high-performance system (more main memory & faster processors & more efficient algorithms)
- **R3:** limiting the number of simultaneously allowed users to one fourth of the number of students
- **R4:** monitoring and automatic restart of the server

- **R5:** maintenance activities are done exclusively in the mornings between 7 and 9 am
- **R6:** user errors are caught and do not lead to system failure
- **R7:** backup every three days

We took care that these countermeasures are less fuzzy than those in experiment 1, and some are even measurable.

In experiment 2, we dealt with an existing system which is known to and used by the participants. Here, it was easiest to define the status quo to be the reference system. In such a reference system, estimations of the varied risk are to be performed differently for countermeasures which are implemented in the reference system and for those which are not. For those countermeasures not realized in the reference system, the varied risk is estimated by estimating the misuse case risk in a system which differs from the reference system by this one countermeasure being implemented additionally.

Material: In this experiment, two methods were applied: method 1 and method 2. During the experiment, the moderator entered the values estimated by the group into spreadsheets which were projected to the wall, so that they were visible for all participants. These spreadsheets calculated risks and benefits automatically.

For method 2, the reference system this time was defined to be the actual system. Some of the countermeasures are realized in the reference system, and for each of these the varied risk is estimated, imagining that this countermeasure was not realized. For those countermeasures, which are not realized in the reference system, the varied risk is estimated assuming that this countermeasure was additionally realized. The difference between these two risks for each misuse case is the benefit achieved by each countermeasure with respect to this misuse case.

Additionally to performing the estimations in the group, the participants individually answered to questions of three questionnaires:

- Questionnaire Q1 evaluates method 1 and was filled out immediately after the execution of method 1. It asked:
 - Question 1.1: How certain are you that the priorities resulting from the group discussion are realistic? (variable c)
 - 1.2: Were there requirements for which you are especially uncertain that they are classified right? If yes: Which? (the list of requirements was offered here) Why? Which information was missing? (variable f)
 - 1.3: Do you think that you have been involved adequately during the discussions and that your proposals and objections have been considered sufficiently?
 - 1.4 Comments (free-text)
- Questionnaire Q2 evaluates method 2 and was filled out immediately after the execution of method 2.
 - 2.1: How certain are you that the probability and damage estimations resulting from the group discussion are realistic? (variable c)
 - 2.2: How certain are you that the priorities resulting from the group discussion are realistic? (variable c)
 - 2.3: What do you think how uncertain are the group's estimations of the probabilities? (If an estimation $p = 2$ times per month presumably lies between 1.5 and 2.5 times, then the accuracy is " ± 0.5 times".) (variable e)
 - 2.4: Were there misuse cases for which you are especially uncertain? If yes: Which? (the list of misuse cases was offered here) (variable f)
 - 2.5: What do you think how uncertain are the group's estimations of the damages? (If an estimation of 4 hours presumably lies between 2 and 4 hours, then the accuracy is " ± 2 hours".) (variable e)
 - 2.6: Were there misuse cases for which you are especially uncertain? If yes: Which? (the list of misuse cases was offered here) (variable f)

- 2.7: Do you think that you have been involved adequately during the discussions and that your proposals and objections have been considered sufficiently?
- 2.8 Comments
- Questionnaire Q3 compared method 1 with method 2 and was filled out after the execution of both methods and after Q1 and Q2.
 - 3.1: Which of the methods were easy to execute and which was difficult? (variable b)
 - 3.2: Explanations and comments
 - 3.3: Which of the methods would you presumably have found easy or difficult, if you had executed it alone?
 - 3.4: Explanations and comments
 - 3.5: How have the group discussion been useful for the results?
 - 3.6: Did this discussion also have disadvantages?
 - 3.7: Which were the advantages and disadvantages of the quantitative risk estimation compared to the intuitive ranking of the requirements?
 - 3.8: Compare the priorities resulting from both methods. Do they reflect your view? (variable d)
 - 3.9: Explanations and comments

Variable g was not tested in experiment 2, because we did not expect any additional information from this quality criterion.

Great care was given to the wording of the misuse cases and countermeasures, the design and content of the experiment material and the instruction of the four moderators. The task of the moderators was to organize the group discussions, to type in the estimations into the spreadsheets, to control the time and to avoid misunderstandings concerning the misuse cases, countermeasures, the definitions of probability and damage and the use of the method. The moderators were not allowed to propose any values. They guided the process by asking questions.

There were two preparatory sessions with the moderators: in the first, the concept and material of the experiment were discussed, in the second, a test run was executed where the future moderators were estimators and the author of the method was the moderator.

5.2 Experiment No. 2: Execution

At the beginning of the estimation workshops, the students got an introduction of twenty minutes to requirements prioritization in general and the two methods, to risk estimation and the significance of the reference system. The execution of the experiment and the meaning of the misuse cases and countermeasures were explained. It was also defined that estimation of probabilities refer to “number per month” in a normal month during lecture time and that damages are estimated in calendar hours lost in a student project (not working hours: for instance, if Sysiphus breaks down at midnight and is re-started at 9 o’clock the next morning, then nine hours are lost in which no one could work, although the work time lost may only be 20 minutes in all). Consequently, risk describes the calendar hours which are lost per month due to a misuse case. Immediately before executing method 2, another five-minute introduction to requirements prioritization was given to each group. Based on our experience from the test run, the agenda allowed 20 minutes for method 1 and 40 minutes for method 2.

This time, the order in which the two methods were executed, was varied. The first four groups started with method 1 and then executed method 2, the other three groups proceeded in the reversed order.

Questionnaires 1-3 were used to ask the participants about their opinion about the methods. Immediately after each method, a questionnaire asked to evaluate this method, and at the end of the experiment, questionnaire Q3 requested to compare the two methods.

After the experiment the discussion moderators were asked about the process of the group discussions and these observations were evaluated qualitatively.

5.3 Influencing Factors in Experiment 1 and 2

In both experiments method 1 and method 2 were executed, but in the second experiment there was no method 3. In the first experiment, 9 countermeasures and 12 misuse cases were treated, in the second experiment there were 7 countermeasures and 6 misuse cases.

In addition to the number of countermeasures and misuse cases, the factors presented in Table 1 had been varied between the experiments. This was done taking into account what we had learned from the first experiment, with the objective to improve the risk estimation in the second experiment. Some of these factors are explained here:

- **Group discussions:** In experiment 1, each participant did estimations individually on a paper questionnaire, in experiment 2, groups of three or four persons discussed jointly until they agreed on one value.
- **Transparency:** In experiment 1, with method 2 the participants estimated probabilities and damages of misuse cases, but did not know which misuse cases risks, countermeasure benefits and countermeasure priorities would result from these estimations. In experiment 2, the estimated probabilities and damages were put into a spreadsheet which automatically calculated these values immediately and which were visible to all group members. It was technically possible to test different values and to check their influence on the result, but was not done often due to time limits.
- **Prioritization criteria:** For method 1, in experiment 1 the participants were asked to rank the countermeasures according to their benefit. When asked which criteria they had used, a large variety of criteria was named. Not two participants shared the same criteria, like cost of changing the implementation, specific need during start phase of the online flea market, competitiveness, cost of not-implementation, or quality of service. Therefore, in experiment 2, we defined more clearly that the countermeasures had to be ranked according to their benefit with respect to a misuse case.
- **For method 2,** in the first experiments we preferred estimating damages in relative values in %, because relative values had been found to be easier to estimate in former experiments by other authors. In the second experiment, we measured damages in lost calendar hours. This was more tangible and could be deduced from the participants' experiment. Probabilities were estimated in % in experiment 1, whereas in experiment 2, they were measured by the number of times per month. This metric was also more tangible und less ambiguous than a probability in percent. A percentage can mean something else for each misuse case.

Table 1. Influencing factors varied in the experiments

Factor	Experiment 1		Experiment 2	
	Method 1	Method 2 & 3	Method 1	Method 2
Number of participants	10	10	24	24

Number of resulting date sets	10	10	7	7
Number of requirements	9	9	7	7
Example system	Ficti-tious	Ficti-tious	Real	Real
Group discussion & moderation	No	No	Yes	Yes
Transpa- rency	Yes	No	Yes	Yes
Order of execution	1 st	2 nd & 3 rd	1 st or 2 nd	2 nd or 1 st
Prioriti- zation criteria	Benefit	Relative damages in %, probabilities in %	Benefit with respect to misuse case	damages in lost calendar hours, probabilities in times per month

6. Results and Lessons Learned

In this section, the results of experiment 1 and 2 are presented and analysed together. These results include quantitative data (time need and variables a-g, summarized in the annex), the results of the data analysis and lessons learned as concluded from the free-text answers in the questionnaires, from discussions with the participants and from observations by the moderators. We here analyze how the factors summarized in Table 1 influence the variables defined in section 3 and the qualitative results of the experiments. Where averages from different samples are compared, we tested the statistical significance of the difference by testing the hypothesis that the expectation values of both samples are equal. If this hypothesis can be accepted with a certainty of 90% or more, we assume that the difference is not statistically significant.

6.1 Influence of Statistics

We expected the following influence of providing statistics about frequencies and damages of specific misuse cases. When the participants are given several statistics:

- This significantly influences the value of the estimated probabilities of misuse cases,
- the standard deviation (relative to the average) of their estimations is lower (variable a),
- they feel more confident about their estimations (variable f).

In experiment 1, the participants were provided four publicly available statistics about security incident frequencies and were asked to re-estimate the probabilities of two security misuse cases. The results are presented in annex B. As can be seen from these results, providing statistics to the estimators did (statistically significantly) influence the value of their probability estimations and also lowered the relative standard deviation (variable a). However the participants were not sure whether the statistics really facilitated their estimations (variable f).

Half of the participants in their free-text comments were still sceptical whether the estimated values based on statistics are more exact than those estimated without. Similar doubts had been uttered by Xie et al. (Xie, 2004) (see section 2.1).

6.2 Improvements Between Experiment 1 and 2

From our experience in experiment 1 we learned more about the challenges and needs of risk estimation and therefore varied some influencing factors in order to improve the results. For example, that instead of a fictitious system, a real system known to the estimators was chosen (because the estimators felt very uncertain about their risk estimations), and estimations were not given individually but in moderated group discussions (to avoid misunderstandings about the method and to bring different experiences and knowledge into the estimation), also the prioritization criteria were defined more clearly and tangibly, and transparency was now introduced into the risk estimation by tool support.

All these changes were expected to improve the results of the experiment with respect to at least some of the criteria defined above, what in fact they did. In this section we present these improvements. In the following sections, we discuss which of the improvements is due to which factor. As so far we have performed only two experiments, the effects cannot easily be attributed to one factor and most of them in fact have more than one cause. However when for instance one improvement is more pronounced for method 2 than for method 1, this is a hint that transparency plays a major role here because transparency was improved for method 2, but was the same for method 1 in both experiments.

The **standard deviations** of the priorities for the same countermeasure but different estimations (variable a, see Table 4 and Table 5) were compared. In Table 6, the standard deviations are normalized, i.e. divided by the average priority, which is equal to $(n+1)/2$ when n is the number of countermeasures. The normalized standard deviations – also called coefficient of variation - of the priorities have indeed decreased. This means that some sources of uncertainty and different perception which were caused by the setup of the experiment could be reduced. The difference is significant: For method 1, this reduction was 25% and for method 2, it was 17%. We must remark that the countermeasures and the participants were not the same in the two experiments.

In terms of **ease of use** (variable b, see Table 7), both methods score better in the second experiment. For method 2, this effect is more pronounced. (However, the differences between both experiments are statistically not significant, only by a certainty of 50%.) This perceived higher ease of use – if it exists - can be expected to be a combined effect of several of the factors varied between the experiments.

The participants in the second experiment believed that their results were more **realistic** (variable c, see Table 8). This is true for both methods, but the effect is more pronounced for method 2. It is interesting to remark that when comparing Table 6 to Table 8, one finds that the coefficient of variation is the lower, the more realistic results were expected by the participants. This indicates that the participants can judge which values were uncertain to estimate and where the other participants might obtain similar results. There is only one exception: In experiment 1, they expect method 3 to yield less realistic results than method 1,

although the standard deviation in method 3 is below that of method 1. (However, these differences are not statistically significant due to high standard deviations of these variables among participants.)

In experiment 2, the **uncertainty** of the probability or damage estimations (variable e, see section A.e) were estimated to be much higher than in experiment 1. These results might not be relevant, as in experiment 1, we got very few answers to this question. In section 6.3, we discuss the difference.

Table 9 presents the **frequency** with which the participants named a countermeasure or misuse case as being especially **uncertain** (variable f, see section A.f). These results indicate that there were statistically significantly fewer uncertainties during experiment 2 than during experiment 1.

(Remark: Overall, a countermeasure was named with a frequency of 0.3 in average, for the three countermeasures R3, R5 and R7 which are defined quantitatively, this average is only 0.15. It seems that the metric made the countermeasure easier to judge, but this question should be re-investigated with more than seven countermeasures.)

Variables d and g were not compared in both experiments because variable d was determined qualitatively in experiment 1 and quantitatively in experiment 2 and therefore was not comparable. Variable g was not determined in experiment 2 at all.

Statistically relevant improvements in experiment 2 compared to experiment 1 have been observed with respect to the coefficient of variation, and less misuse cases and countermeasures were named as being difficult to estimate. Statistically not relevant improvements were observed with respect to ease of use and participants' expectation of how realistic the results are,

6.3 Influence of Moderated Group Discussion

The group discussions compared to individual estimations were expected to not only demand more time, but also to improve the quality of the results and the participants' perception of the process. The results of a group should be better than those of each individual, because during the discussion the knowledge of several persons is added, misunderstandings with respect to the method or the case are discovered, missing information investigated or common assumptions are made. In addition to comparing variables a-f in both experiments, we ask the participants explicitly how they perceive the discussions.

One very strong influence was observed on **time need**. We expect the time need (see Table 3) to be proportional to the number of estimations to perform (which differs a lot between the methods) and to depend on whether the estimations were done alone or discussed in a group. In method 1, the estimation of one value with 3-4 participants took 3.4 times as much time as individual estimations. In method 2, this factor was 2.3. (All these ratios are statistically significant.) While the group discussions took more time than individual estimations, the group discussions presumably contributed to the improvements discussed in section 6.2. Whether and to which degree this is true, should be investigated in additional experiments.

In experiment 2, the **uncertainty** of the probability or damage estimations (variable e, see section A.e) were estimated to be much higher than in experiment 1. This is difficult to explain. One would expect that the values were more certain in experiment 2 due to the group discussions and the other factors varied between both experiments. It is not sure whether these differences are significant, because we got very few data for this variable from experiment 1. If in fact after group discussions participants expect higher uncertainties, we propose two possible reasons for this:

1. The discussion of alternative risk scenarios and differing experiences in experiment 2 revealed the variety of perspectives and experiences, and consequently irritated the participants concerning the interval in which the value really lies. In the individual

estimations in experiment 1, no such irritation took place. Each participant could feel more certain about his/ her opinion.

2. In experiment 1, the participants were asked to estimate the relative uncertainty in percent, while in experiment they were asked to estimate in absolute values. Therefore, the two values might not be comparable.

We wondered whether all participants perceived the same countermeasures or misuse cases as **difficult to estimate** (variable f, see section A.f). And whether in experiment 2 the members of a group share this perception. Such effects might have been caused by discussions in the groups which having focused on the same misuse case for a long time.. Some single countermeasures or misuse cases were especially certain or uncertain for many participants, but no evident correlation was visible among the opinions of the members of the same group.

The other variables observed allow no conclusions specific to group discussion. Yet the estimators were asked some qualitative questions with respect to group discussion and the replies are discussed in the following.

In experiment 2, we asked: Do you think that you have been **involved adequately** in the discussions and that your proposals and objections have been considered sufficiently? Concerning method 1 (question 1.3), an average of 1.58 points was attained (from an interval of [-2, +2]) and concerning method 2 (question 2.7) it was 1.25 points. This difference is statistically significant.

We asked the participants explicitly how they perceived the group discussions in questions 3.3 to 3.6. We were interested in the qualitative content of their replies, but also whether more advantages or disadvantages were named. In questions 3.5 and 3.6, advantages and disadvantages of the group discussions were named: The participants named more advantages than disadvantages (20 advantages versus 6 disadvantages). As advantages they repeatedly named: discussion led to a better understanding of the meaning of countermeasures, misuse cases and the method, clarification of misunderstandings, taking into account different experience (by calculating averages, by eliminating exceptional experiences, by discussing new aspects and solutions). Disadvantages were: individual opinions are neglected, different roles of the participants influence the group consensus (here: project managers dominated), higher time need.

In free-text comments, they expected that estimations would have been difficult when done alone, because knowledge exchange was useful (named twice), one participant emphasized that moderation was important, one felt that estimations would have been easier because no discussion would have taken place, one would have demanded a clearer specification of the countermeasures, and three expected no difference.

In question 3.1, we had asked which of the methods was easy to use and which was difficult. In 3.3, we asked how difficult the methods would presumably have been, if executed alone. We compared the answers to question 3.3 with 3.1. The participants expected that doing the estimations alone would have been somewhat more difficult or less easy than it was during the experiment. For method 1, 1.00 instead of 1.13 points; method 2, -1.13 instead of -0.92 points. The difference is statistically significant for method 1, but not for method 2 because a higher variance of the data.

The moderators observed that the student who had the role of project manager within the team in almost all the groups organized the finding of the consensus.. This was probably not only due to the project manager role, but also because the project managers had more experience with the example software. They had started using Sysiphus several weeks before the other students and most of the project managers also knew the software from earlier courses.

The effect of moderated group discussions compared to individual estimations was difficult to isolate quantitatively in these two experiments. The participants' answers indicate that they

perceive the group discussion as positive. However, the group discussions demanded more time and roles influence the discussion.

6.4 Influence of Transparency

Transparency here means that when doing the probability and damage estimations, the participants can see (and control) their influence on the resulting countermeasures priorities. In experiment 1, method 2 was intransparent, while in experiment 2 we performed it in a transparent way. In experiment 2, the estimated probabilities and damages were put in by the moderator in a spreadsheet which automatically calculated these values immediately and which were visible to all group members. It was technically possible to test different values and to check their influence on the result, but this was not done often due to time limits. Method 1 was transparent in both experiments. We expect that such a transparency leads to corrections of the estimations during a plausibility check of the priorities. We also expected the participants to perceive the quality of the resulting estimations and priorities as better and to feel a higher trust in the method.

In experiment 1, we tested whether “blind” estimations of probability and damage which are done without knowing their effect on the resulting priorities lead to good results. However, in Q5a, 9 out of 10 participants marked “because the estimations were split up in single steps and the result of one’s estimation is not predictable” as a reason why his/ her results were so different with different methods. In the subsequent experiment 2, we hoped that transparency would reduce the inconsistency problems observed in experiment 1, where 6 out of the 10 participants obtained at least one negative countermeasure benefit. Such a negative value signifies that the implementation of a countermeasure did not lead to a risk reduction but to a risk augmentation. In fact, no such errors were observed in experiment 2, because they could easily be detected.

Transparency, as was expected, was observed to lead to corrections of the estimations during experiment 2. The participants could compare the resulting risks to those of other misuse cases and checked for plausibility. Due to time constraints in the experiment, this was done in only few cases, but sometimes the probability or damage estimations were corrected.

As has been discussed above, there have been several improvements observed in experiment 2, compared to experiment 1. If this effect is more pronounced for method 2 than for method 1, this can be a hint that the higher transparency causes part of this improvement, because for method 1, the transparency was the same during both experiments. The improvement in fact was higher for the **ease of use** (variable b, see Table 7: For method 1, the improvement was 0.12 points, for method 2 it was 0.28.) and the results are expected to be more **realistic** (variable c, see Table 8, where method 1 gets a plus of 0.04 more points, but method 2 of 0.07). The statistical significance of these differences for variables b and c however is low (50% certainty).

However, transparency seems to have no major effect on the **standard deviations** (variable a) the improvement was more pronounced with method 1 than with method 2.

6.5 Prioritization Criteria and Metrics

In experiment 1 for method 1 the results of different persons differ widely from each other. One reason for this might be that method 1 here did not define clear prioritization criteria, but asked the participants to rank the countermeasures concerning their benefit. When being

asked about their criteria, each of the participants named different criteria, usually quality attributes like security or usability, but also user benefit or administrator benefit, risk or to surpass competitors. We hoped that using clearly defined prioritization criteria in method 1 would lower the standard deviation of the results. We therefore defined in experiment 2 that the countermeasures were to be ranked according to their benefit relative to a specified misuse case.

In methods 2 and 3 we had clear criteria (probability and damage) and nevertheless the results of the participants differed a lot. In experiment 2, more tangible measures were used (explained in section 5.3). We hoped to reduce the influence of misunderstandings and consequently the standard deviation of the resulting priorities.

Indeed such a reduction of the **standard deviations** in experiment 2 was observed for methods 1 and 2. For method 1, this effect was stronger than for method 2.

Table 9 presents the **frequency** with which the participants named a countermeasure or misuse case as being especially **uncertain** (variable f, see section A.f). In experiment 1, damage estimation has been statistically significantly more difficult (in method 2) than probability estimations, whereas in experiment 2, no statistically significant difference is found. This can be an effect provoked by the different metrics used in the two experiments.

The prioritization criteria, damage metrics and probability measures have to be chosen carefully. There were hints that the risks referring to measurable countermeasures are easier to estimate.

6.6 Order Effects and Learning Effects

In experiment 2, we tested order effects: Four groups (number 1-4; 13 participants) executed method 1 first and then method 2, the other three groups (number 5-7; 11 participants) in a second shift (supported by the same moderators) proceeded vice versa. There were differences observed between the results of groups 1-4 and 5-7.

The method was executed second in order **took less time**. From the data in experiment 2, we estimate that about 12% of the time is needed for general explanations and clarifications.

Differences were also found with respect to the variables c, d, e and f:

Variable c: The participants of groups 5-7 expected their estimations to be more **realistic** than groups 1-4 did: Concerning method 1, groups 1-4 gave an average of 0.85 points versus 1.27 given by groups 5-7. With respect to method 2, the probability and damage estimations were expected to be realistic with -0.23 versus 0.55 points, and the resulting priorities with -0.08 versus 0.45. This means that groups 5-7 considered their results to be more realistic in general, for both methods. All these differences have been found to be statistically significant.

The findings are similar for accuracy (variable d, question 3.8): Method 1 received 1.23 versus 1.36 points, and method 2 0.00 versus 0.27. For method 1, the difference is statistically not significant, but for method 2 it is.

The estimated uncertainties of the probability estimations (variable e, question 2.3) were 41% versus 28%. When being asked for misuse cases which were especially uncertain (question 2.4), groups 1-4 named one misuse case (variable f) with a frequency of 0.23 and groups 5-7 only with 0.18. With respect to damage estimation, these frequencies were 0.19 (groups 1-4) versus 0.27 (groups 5-7) (questions 2.5 and 2.6). With respect to method 1, the frequency in groups 1-4 was higher: 0.20 versus 0.09. These differences are statistically significant. Each participant of group 1-4 named at least one countermeasure here, but in groups 5-7 only 5 out of 11 participants did.

In summary, groups 5-7 seem to have felt more confident about their results (but not about the ease of use of the method and about damage estimation), even for method 2 which they executed first. In this experiment setting, this probably signifies a learning effect of the

moderators. In experiment 2 we proceeded in two shifts for practical reasons: Groups 1-4 started and immediately afterwards, the same moderators executed the same experiment with groups 5-7. It is possible that during their second run, the moderators could answer to questions better, explain the method better and in general felt more confident, which influenced the participants' perception

6.7 Comparison of the Methods

In both experiments, methods 1, 2 and 3 were compared with each other with respect to time need and the variables a-g. We also cite some remarks of the participants about the differences which they experienced between the methods.

Time need: As expected, method 1 was fastest. This is not only because there were less estimations to do per countermeasure, but also the time need per estimation was lower. In the prioritizing of the same (number of) countermeasures, in experiment 1, the relation between the total time need was 1 : 4.8 : 6.5 (for methods 1, 2 and 3) and in experiment 2, it was 1 : 2.1 (for method 1 versus 2) (see Table 2).

The time need per estimation in experiment 1 was 1 : 1.6 : 1.5 (for methods 1, 2 and 3) and in experiment 2, it was 1 : 1.1 (for method 1 versus 2) (see Table 3). This shows that the higher time need mainly results from a higher number of estimations to do and to a lesser part (but statistically significantly) from the fact that damage and probability estimations are more difficult than the decisions in method 1.

In experiment 1, method 3 needed 38% more time than method 2 for prioritizing the same (number of) countermeasures (see Table 2). It is interesting to remark that with method 3 the time need per estimation was lower than with method 2. This indicates that the step-by-step procedure in method 3 facilitated damage estimations. The same conclusion is implied when in experiment 1 (Q5b), 7 out of 10 participants wrote that the relative step-by-step damage estimation in method 3 was easier than the absolute value estimation in 2. (Two persons thought that it was easier in method 2, and one said that both were equal.)

The **standard deviation** (variable a), see Table 4 and Table 5: The differences of the standard deviations of the methods in the same experiment are very low and statistically not significant. However there seems to be the highest consistency of the resulting priorities in method 3, whereas method 1 seems to be slightly better than method 2. One might wonder whether a higher consistency of the priorities means that they are more realistic. It is possible that in method 3, the estimators were simply more strongly guided towards similar results. We do believe that the risk estimation in principle predefines the estimator's perspective by the misuse case definition. In method 3 the linking of misuse cases to business goals seems to do guide the estimator even more strongly. This can be an advantage if the business goals are properly defined, but can also be a disadvantage because other perspectives which are not considered by these business goals might be underrepresented by the countermeasure priorities.

Ease of use (variable b), see Table 7: In both experiments, method 1 was judged to be easier to use than method 2 and 3. Method 1 was found rather easy to use, method 2 rather difficult. It is not clear whether method 3 is really easier to use than method 2, although method 3 gets much more points than method 2. This is because in the experiment in method 3 we reused the probability estimations of method 2. Even if the damage estimation in method 3 was experienced to be easier in method (Q5b, see above), the probability estimation is not different. So, we cannot be sure whether the method 3 would seem easier to use if it included the probability estimations. The difference found between method 1 and method 2, however, is statistically significant.

Realistic priorities (variable c), see Table 8: The participants expected method 1 to deliver the most realistic priorities (“rather realistic”). Method 3 got more points than method 2. The difference between method 1 and 2 is statistically significant.

Accuracy (variable d), see annex A.d: We summarize the qualitative answers obtained from experiment 1 by saying that intuitively the participants felt that method 1 reflected their subjective priorities best, while they thought that method 2 and 3 must deliver more objective and therefore better results. Methods 2 and 3 were not clearly perceived as being more accurate. Methods 2 and 3 were not explicitly compared to each other. In experiment 2 (questions 3.8 und 3.9), method 1 again rated better than method 2.

Variable e (**uncertainty** of the estimated values) makes no sense with respect to method 1 and therefore is not considered here. In Table 9, we see that in both experiments, the risk estimations in method 2 and 3 led to a higher number of estimations where the participants felt especially uncertain (variable f). Participants felt clearly more certain about their estimations (variable g, see Table 10) with method 1 than with the others, whereas method 2 and 3 were rated almost equally.

In both experiments, the risk estimations in method 2 led to a (statistically significant) higher number of estimations where the participants felt especially uncertain (variable f).

Which method is best? Method 1 rated better than methods 2 and 3 with respect to time need and the quality variables b-c and e-f. Only with respect to variable a this was not the case. Although the differences of some of these variables are not statistically significant when considered alone, the fact that almost all quality criteria show the superiority of method 1 (ranking) to the risk estimations (methods 2 and 3) is a clear result. One might now wonder whether this is due to the experiment setting, where the misuse cases and countermeasures were defined by the experiment organizers and the participants had no previous experience with risk estimation. We will discuss threats to validity later on.

In experiment 2, the participants were asked which advantages and disadvantages the quantitative estimation in method 2 had, compared to the intuitive ranking in method 1 (question 3.7, free-text field). Among 19 free-text replies, 13 mentioned advantages and 18 disadvantages. From this we conclude that the quantitative estimation was experienced as rather being inferior. As advantages the participants named: objective measure, own experiences can be contributed, schema, order of magnitude, more details. As disadvantages they saw: difficult estimation, especially when information is missing, high time need for the same result, dependency on many factors, uncertainty of the estimated values, coming to a consensus is more difficult. The latter can be explained by the fact that in method 1, there are only few values to choose from, compared to method 2.

Whether method 3 is better than 2 could not clearly be decided. We found slightly better results for method 3, but they were not statistically significant.

The step-by-step estimations of method 3 were considered to be an advantage by several participants because they ask simpler questions which can easily be answered. On the other hand, these step-by-step estimations amplify estimation uncertainties. We assume that probability and damage estimations are uncertain by $\pm x$ percent points. In method 2, the benefit of a countermeasure is $p_{\text{var}} \cdot d_{\text{var}} - p_{\text{ref}} \cdot d_{\text{ref}}$. When each of these values is uncertain by $\pm x\%$, then each risk is uncertain by $\pm 2x\%$ and the benefit of a countermeasure by $\pm \sqrt{2} \cdot 2x\%$. In method 3, the uncertainty of the countermeasure benefit is much higher: For the reference risk, the probability is uncertain by $\pm x\%$, but for the damage $\pm 4x\%$ because of the cascaded multiplication of the estimations down from the business goal. For the varied risk, the probability is uncertain by $\pm x\%$, but for the damage even $\pm 5x\%$. This means that the varied risk is uncertain by $\pm 6x\%$ and the reference risk by $\pm 5x\%$, and their difference, i.e. the benefit of the countermeasure by $\pm 8x\%$!

In experiment 1, $x\%$ is 10%, in experiment 2 even 40%. Even when assuming the optimistic 10%, the benefit estimations resulting from method 2 are uncertain by about 30%

and of method 3 by 80%. As the benefit values can be quite far apart and for ordering the countermeasures according to their priorities, the absolute values are less important than the right relations, we had a look at such differences to find out whether 30% respectively 80% are high. In experiment 1, we analyzed the results of four participants and calculated the relative differences between benefits of countermeasures ranked with consecutive priorities i and $i+1$. For method 2, they varied between 14% and 530%. 25 out of 32 (= 78%) of these differences were 30% and more. But for method 3, they varied between 8% and 380%. Only 11 out of 32 = 34% (mostly at the top and the bottom of the hierarchy) of these differences were 80% and more. This means that for method 2, the uncertainty of 30% does not necessarily mean that the countermeasures priorities are unreliable, because the differences between the estimated benefits were large enough. But for method 3, the high uncertainty is a problem. So, for using method 3, more reliable estimates are needed. The high uncertainties in method 3, due to the consecutive multiplication of estimated values, we saw as a major disadvantage compared to method 2. This is why method 2 was not tested again in experiment 2.

6.8 Further Lessons Learned

Additionally to the results discussed above, we gathered lessons learned from free text feedback of the participants in the questionnaires and in discussions, as well as from the observations of the discussion moderators.

Risk estimation turned out to be difficult, as was expected. The resulting priorities of the participants in all methods differed a lot. The participants themselves (in experiment 1, Q5f) proposed the following reasons for these deviations (the first three were named by more than one person):

- Different criteria and goals of different estimators
- Different experiences
- Uncertainty of the estimations
- The end result (the priorities) was difficult to foresee for the participants
- Missing information led to differing assumptions
- Missing experience and technical competence
- Misunderstandings concerning the method
- Missing knowledge about market and reality
- No feedback about the other participants' estimations, which might have led to more realistic values
- Time pressure

To estimate risks reliably and to feel certain about one's estimations, one needs a lot of information about the system, the usage, the environment. According to participant answers in the questionnaires and according to our observations during both experiments, such information is:

- To estimate the reference risk, personal experience with the system is useful.
- The varied risk is estimated on the basis of "What-if" questions. For these estimations, practical experience is required with countermeasures, which have not implemented so far, in order to come to realistic expectations about their effect. It has to be clearly defined what the system would be like, if a countermeasure was not implemented or implemented additionally.

- Expert knowledge is needed, e.g. management knowledge (about business goals), user knowledge (e.g., from the user perspective about frequent misuse cases and damages caused), and technical knowledge (e.g., from the technical perspective about frequent misuse cases and damages caused, as well as technical possibilities to mitigate misuse cases). An example of where the lack of knowledge caused difficulties, is: In experiment 2, the participants from their user experience knew how often a user observes a system failure. However, each of three reasons given referred to another misuse case (because each demands specific countermeasures), and the participants could only guess the relative occurrence probabilities of these misuse cases. To have such expert knowledge available is difficult even if the system exists, is well known to the estimators and is used regularly.
- knowledge about the future and the dependency of risks

In Q5a of experiment 1, 3 out of 10 participants marked “missing information” as a reason why his/ her results from different methods were so different and 4 marked “misunderstandings”. Information which the participants regarded as missing for instance was: the number of system users, the cost for setting up a support centre, how well are the servers of the company protected by measures other than encryption, the qualification of the personnel, the knowledge of the users. Xie et al. (Xie, 2004) had also found that risk is highly project and company specific. We wonder whether it would have been practically possible to read and understand all information necessary for a good estimation, during an experiment. We expect that it makes only little sense to estimate risks for a fictitious system during an experiment. Method evaluations must take place in a real project, or at least in a real student project. The persons involved will have a lot of this information available from their experience with their software, the environment etc. In fact, in experiment 2, the results were better, but it cannot be quantified how much of this improvement is due to the fact that an existing system was used in this experiment

The participants of a prioritization workshop only need half an hour of **training** to understand the principle, but they need **tool support** and a **moderator** who guides them through the estimations step by step. The prioritization workshop has to be prepared by performing a MOQARE analysis. Because of the n-m-relationships between misuse cases and countermeasures the benefit calculations are not straightforward and spreadsheets supporting them cannot be reused without adaptation.

The **moderator** needs a good knowledge of the method and much experience to be able to answer all questions and to guide the group well.. He/ she should ideally have a perfect understanding of the theoretical background of the method and a thorough preparation. This preparation includes the decision on which misuse cases to consider, on how to write them down (phrasing), on which statistics and other information to supply.

In general, the participants all expressed that they did not feel sure about their probability and damage estimations. During the experiment, a discussion arose, as 20% did not mean the same for all misuse cases, there could be no general and satisfactory rule given for the probability estimation. In experiment 1 (Q5a), 8 out of 10 participants marked “because risk estimations are difficult in general” as a reason why his/ her results from different methods were so different.

While we could provide **statistics** about security incidents, this was not the case for other quality attributes. In these fields risk estimations and therefore also the recording of misuse probabilities and damages are not done as systematically as for security misuses (example: misuse case “user error prohibits sale”). Such statistics could support risk estimation and requirements prioritization a lot. Statistics about probabilities and damages caused must be available, relating to a system and environment as similar as possible to the present one, if possible with incident statistics or experiences from the same company. As public statistics

rarely apply to the same environment as the system under consideration, there will still be high uncertainties in the results of the estimators due to adaptation.

Granularity of the misuse cases: In section 2.1, we described that grouping misuse cases and countermeasures is a means of taking into account dependencies among them; it is also a means of saving time. The moderators observed that when misuse cases and countermeasures are too general (e.g. including a whole group of misuse scenarios), they are difficult to estimate because we have to average over many scenarios; when they are too detailed, the time need is increased and dependencies among these fine-grained misuse cases irritate the estimators. In free-text comments, the participants also criticized that the misuse cases were too general.

Sometimes the group value is calculated as the **average** of several estimations. This is the case when group consensus on one value could not be achieved or when several scenarios had to be considered for one misuse case. Then, the result differs whether the average is made for probabilities and damages separately or for the risks (and then the probability and damage estimation are concluded on). As in the following simple example: one estimation of risk is $1 \times 1 = 1$ and the other $2 \times 3 = 6$ -> the average probability is 1.5, the average damage 2 -> the risk is 3. But the average of the two risks is 3.5.) Such a difference was observed once during experiment 2, where different types of misuse cases were grouped in one, and consequently probability and damage were correlated. There was one frequent misuse case with low damage and another one which happens rarely, but causes high damage. In this case, the average should not be calculated from probability and damage estimations but from the risk.

It is difficult to estimate the damage of all misuse cases in the same **measure**, because in practice, damages influence different goals and therefore are measured in different units like Euro, calendar time, work time, score received for the homework. Misuse case risks are difficult to compare. Maybe it is even impossible to measure benefit and damage with one and the same measure. Instead, one could choose points as a unit (like in FMEA (Stamatis, 2003) and many other approaches).

We expect that the results of the estimations are sensitive to the definitions and the **wording**. In a prioritization workshop, the estimators should agree on the definitions and wording and rephrase countermeasures and misuse cases if necessary. This signifies an additional coordination effort for them. As a positive side effect, these discussions lead to the quality assurance of the requirements. However, in the experiment setting, wording could not be modified because this would have endangered the comparability of the estimation results.

The misuse cases define the **perspective of the estimator** (e.g.: user perspective, developer or maintainer perspective). This can have advantages as well as disadvantages. A clearly defined perspective helps the estimators to implicitly consider the business goals underlying the countermeasures like user satisfaction or low maintenance cost. However, it is difficult to estimate from an unfamiliar perspective.

Some participants felt that the benefit of a countermeasure **with respect to one misuse case** does not measure its benefit for the whole system. They also criticized that risks and disadvantages caused by a countermeasure are not taken into account by the method. (Remark: In practice, the estimators should define new misuse cases, if they discover important misuse cases caused by the implementation of a countermeasure and treat these misuse cases like the others.)

Some lessons learned refer to the **experimentation**. In the experiments, estimators needed and welcomed clear rules and step-by-step instructions for each misuse case (e.g.: “Imagine that the misuse case happens 10 times and calculate the average damage.”) and they demanded unambiguous wording and definitions of the countermeasures and misuse cases,

ideally in a quantified way. The results of the estimations are expected to be sensitive to these definitions, to the wording, and to the granularity of the misuse cases and countermeasures. This means a high preparatory effort for an estimation experiment. In experiment 2, the text for the instructions was 3 to 4 times the volume than the misuse case and countermeasure descriptions. (These long instructions considered all ambiguities and misunderstandings which occurred during the double pre-test.)

The need of support in the risk estimation by **clear rules** must be emphasized. While in the experiment these rules had to be defined by the moderators in order to produce comparable results, in practice, rules, definitions and assumptions can be defined by the estimators themselves, but should be documented.

6.9 Discussion of Validity

The validity of an experiment means that the experiment measures exactly what should be measured.

Wohlin et al. [WRH02], following the classification of Cook and Campbell (Cook, 1979), distinguish between conclusion validity, internal validity, construct validity, and external validity. Conclusion validity is concerned with the relationship of the treatment and the outcome, i.e. whether there is a statistically significant relationship. Our low sample sizes (low number of participants as well as low number of requirements prioritized) are an issue here. The low numbers were problematic in hypothesis tests, because many effects observed were not statistically significant. However, the numbers of requirements and participants were limited by practical restrictions. Nevertheless, we chose to perform the experiments, because scalability of the methods was not a topic under investigation. The experiments described here were intended as preliminary investigation, the experiment effort had to be manageable for practical reasons, and the numbers are not lower than in comparable investigations (see section 2.2). Because many observed relationships were not statistically significant, this publication mainly offers indications on relationships, but no proofs. We believe that the results give useful hints for future experiments and applications of the method in practice.

We also believe that many practical challenges observed during the experiment would not have happened in a real project. In the experiment, the moderator had to define the wording of the countermeasures and misuse cases to obtain comparable results in all groups, while in practice, the estimators themselves would define them in a way which seems optimal for them.

As we performed only two experiments so far and experiences from the first experiment were used to improve the execution of experiment 2, several variables differed in these two experiments. Therefore, the observed effects can be due to several factors and not definitely be attributed to one factor. Further experiments should more reliably test the correlations observed. For instance, for method 1 the differences in the results of experiment 1 and 2 can be attributed to the group decisions or to more clearly defined prioritization criteria. For method 2, the differences lay in group discussions and the transparency of the results. Moreover, not the same number and the same misuse cases and countermeasures were treated in the two experiments; in experiment 1, we used a fictitious example, in experiment 2 an existing system, that was known to the participants. Therefore, further experiments should be made with the same countermeasures, but new combinations of the influencing factors (summarized in Table 1), like transparent individual estimations or group discussions without transparency.

Internal validity is threatened if a relationship is observed between the treatment and the outcome, although there in fact is none. This can happen when the observed effect is caused

by the treatment. Such effects can generally be caused by the order in which the methods are applied. In experiment 1, all participants applied the methods 1, 2 and 3 in the same order. In experiment 2, the order was switched, but there could have been a learning effect on the side of the moderators or any other effect which caused a common difference between those groups who executed method 1 first and then 2 and the others where it was vice versa. However, all groups used the same material, shared the same introductory training and all participants took part on the same afternoon, so no history effect or opinion exchange between participants could have been possible.

Construct validity refers to the extent to which the experiment setting actually reflects the construct under study, e.g. the ability of the measure chosen. To avoid such difficulties, we chose several variables to measure what a “good” requirements prioritization method is. With regard to several effects, we asked open questions and did a qualitative rather than a quantitative test, so the participants could express their opinions freely.

One explanation why we received very different estimations from different participants was, that they might have taken different perspectives when estimating, e.g. the perspective of a user, a developer, maintainer or manager. This variety of perspectives is realistic and could also be found in an industry project team, because in the requirements prioritization these different views should be taken into account. Partly, the perspective of the estimator was predefined by the wording of the misuse case. This is no threat to validity, but part of the method.

External validity is associated with generalization. If there is a causal relationship between the construct of the cause and the effect, can the result of the study be generalized beyond the scope of our study? We have discussed before, that there is a difference between an experiment where the participants prioritize requirements in an artificial example or whether they prioritize requirements in a system which exists, which is in operation and which they know from the user and developer perspective.

The external validity is an important issue in student experiments. To find out whether a method can well be used in the software engineering practice, it should ideally be tested in real projects by practitioners. Nevertheless, for practical reasons, methods regularly are tested by subjects who are students. This is called “convenience sampling”. Robson (Robson, 2002) states that: “Convenience sampling is sometimes used as a cheap and dirty way of doing a sample survey. You do not know whether or not the findings are representative. [...] Nevertheless, studies with students as subjects have made important contributions to empirical software engineering (Carver, 2003).”

To what extent are students representative of real stakeholders in real projects? This is a question regularly discussed and investigated empirically. Some of these studies have found that there are no significant differences compared to professionals, e.g. when estimating the effect of ten factors on time to market (Höst, 2000) or with respect to the improvement observed when using a software engineering process (Runeson, 2003), while others have found that there are significant differences, e.g. (Remus, 1989). “The fact that different studies come up with different results is not very surprising. In some areas it is suitable to use students and in others it is not. However, it is very important to clarify under which circumstances students are useful and not.” (Robson, 2002)

Tichy [Tichy00] gives eight hints for reviewing empirical work. One of these hints is named “Don’t dismiss a paper merely for using students as subjects” where he outlines four different situations where it is acceptable to use students as subjects. These are:

- When the students have been trained well enough to perform the task they are asked for.
- To establish trends: when comparing methods, the trend of the difference if not its magnitude can be expected to be comparable to that of practitioners.
- To eliminate hypotheses: if there is no effect observable in the student experiment, it is very unlikely that an effect is observed with professionals.

- Student experiments as a prerequisite for experiments with professionals.

This means that student experiments are important and helpful for initial studies on a question. Observing trends was a major goal of our experiments.

Tichy especially argues for experiments with computer science (CS) students: “In particular, CS graduate students are so close to professional status that the differences are marginal. If anything, CS graduate students are technically more up to date than the ‘average’ software developer who may not even have a degree in CS. The ‘professional’, on the other hand, may be better prepared in the application area and may have learnt to deal with systems and organizations of larger scale than a student.

Studies have found that mere length of professional experience has little to do with competence. In other words, you can’t use the argument that professionals with years of experience will necessarily solve a given problem better than appropriately prepared (graduate) students. If scale or application experience matters, then the story may be different.“ We are confident that our students did as reliable and realistic estimations as possible, especially in experiment 2, as they are already quite experienced with the system under consideration and also have programming experience in the relevant (university) context.

We believe that in fact professionals would not have had more experience with risk estimation methods than the students. The professionals’ advantage would rather be their higher experience with the system and countermeasures under consideration. In practice, professionals will expectedly ask less for guidance and take a more active part in adapting the method and the wording of the misuse cases to their needs than students can in an experiment. For professionals, the resulting priorities would be more important, while for the students applying the method correctly could be more relevant. However, to professionals methods 2 and 3 would have been as new as to the students. This in fact we wanted to test in the experiments: How usable are these methods for someone with no previous experience. Carver et al. (Carver, 2003) emphasize one difference between students and professionals: In student experiments, a method ‘is being measured in the early stages of the learning curve’. This is true in our experiment. For instance, unlike the experiment participants, the authors of this publication with their experience in risk estimation feel very confident with regard to their own results. Even if the scope of the risk is uncertain, they trust in the relations between the risks.

Although our students said they lack the technical and market knowledge necessary for realistic estimates, such knowledge is not fully available to professionals either (Xie, 2004). First of all, we do not claim that our experimental experience with a fictitious case is equally valid for a real project. In an experiment, the case and the system environment cannot be defined in detail due to practical limitations. However, experiment 2 more realistically simulated the situation in a prioritization workshop in a real IT project team, where the developers have no experience with risk estimation and little practical experience. Nevertheless, we expect different results when experienced estimators perform the same task.

7. Conclusion and Future Work

Tichy [Tichy00]: “The reality of even the most rigorous approach to empirical work is that experiments normally constitute only a small step forward. By their very nature, experiments explore the relationships between a few variables only, while the real world is far more complex. Due to their limited scope, experiments merely gather evidence.“

In the two experiments described above, requirements prioritization based on risk estimation was investigated. Our present experiments highlight challenges of risk estimations

and as an empirical pilot study on this topic, they can serve as a basis for the design of more targeted experiments.

The experiments provide many insights about the challenges and needs of risk estimation. By learning from the feedback of experiment 1, the quality of the results of risk estimation and the participants' trust in the method could be improved in experiment 2. The following lessons learned on risk-based requirements prioritization should be taken into account in future experiments and method developments:

- Group discussions and their moderation are important have positive effects, although group discussions are time-consuming.
- Risk estimation is difficult and requires a lot of information about the system and its environment. Statistical data or own experience about risk probabilities and damages caused are helpful. In an experiment, a real system should be used which the estimators have practical experience with, if possible from different perspectives.
- Providing statistics to the estimators did influence the value of their probability estimations and also lowered the relative standard deviation. However the participants were not sure whether the statistics really facilitated their estimations.
- Transparency is useful, that means to see the effect which each probability and damage estimation has on the resulting risks (and indirectly on the priorities). Tool support which automatically calculates risk can facilitate this transparency. Transparency was found to be advantageous in terms of ease of use and that results are expected to be more realistic, although the results were not statistically significant.
- Prioritization criteria, damage metrics and probability measures should be defined clearly and tangibly. Requirements should be unambiguous and quantified where possible.
- The participants' experience with the method and also the moderator's experience enhances confidence in the results.

These results do not contradict the experiences of other researchers about risk estimation and requirements prioritization, but they are more detailed, what means that we observe the influence of more factors than others and we do this quantitatively. For instance, Feather and Cornford (2003) observe that important for successful risk and benefit estimation are the involvement of experts and "A facilitator is needed to direct these sessions." They support the experts by providing a knowledge base of known misuse cases and countermeasures for the application domain (i.e., spacecraft and software development).

So far, we performed two experiments, but varied several influencing factors: In experiment 2, compared to experiment 1, we performed moderated group discussions instead of individual estimations, we provided transparency of the risk estimations, varied the order of method execution, countermeasures were less fuzzy and more often measurable, prioritization criteria more tangible and clearer, and we used a real system instead of a fictitious one. These variations had statistically significant effects. Some more experiments would help to find out which of the observed effects were caused by the variation of which factor mainly.

Some new questions arose from the observations and quantitative results which should be investigated in more depth. For example the question whether measurable countermeasures are easier to estimate. The prioritization criteria, damage metrics and probability measures must be carefully chosen. There were hints that this might be the case. It is also possible that damage and probability are easier to estimate in points as done in FMEA (Stamatis 2003)

As the estimations of different persons vary a lot, one should test the reliability of the results by repeating the same estimation with the same persons. Learning effects and intermediate discussions among participants however, could bias such a re-test.

Method 1 (the ranking of requirements in two steps) rated better than method 2 (the risk-based prioritization) with respect to time need and almost all quality variables, except for the standard deviation. Although the differences of some of these variables are not statistically

significant when considered alone, the fact that almost all quality criteria show the superiority of method 1 to the risk estimations (method 2) is a clear result. One might wonder whether this superiority of the ranking method over risk-based prioritization is due to the fact that in the experiment the subjects were students. We believe that the participants' feeling of uncertainty during our two experiments and other disadvantages of risk estimation observed can partly be explained to be a beginner's problem of someone with no experience in risk estimation. Some working experience probably is necessary to gain confidence in the method and its results. In fact, after an industry case study using the same method for prioritizing requirements in a real software project, the participants said that the method is easy to use and leads to results which are realistic and useful. Feather et al. (Feather and Cornford 2003, Feather et al. 2006) also use risk estimation, even for high numbers of requirements, successfully. Therefore, we do not conclude from our experiments that risk-based prioritization must be abandoned, but that the participants must be carefully chosen and prepared. The influence of the estimators' expertise can not be underestimated. Feather and Cornford (2003) observe that their "combined expertise" must encompass goals, requirements and constraints, misuse cases, as well as preventative, detecting or alleviating countermeasures.

The high time need and degree of uncertainty of the risk and benefit estimations relativize the usefulness of quantitative risk estimations. There is a saying that in project management it is not the project plan which is important, but the process of planning. Ambler (Ambler, 2002) remarks: "Modeling is similar to planning - most of the value is in the activity of modeling, not in the model itself." We would say that the same is true for risk estimation and requirements prioritization. Despite all challenges met in the experiments, we believe that risk estimation is a good means of discussing priorities of countermeasures. Risk – among other criteria – is an important prioritization criterion. As a side effect, this process forces to phrase the requirements comprehensibly and to identify open questions and missing knowledge, and it requires stakeholders with different experiences to communicate.

As a compromise, we recommend to invest the effort for risk estimation only for the most critical requirements. It probably is most efficient to first prioritize the requirements with a simpler method and then to use risk estimation for analyzing some especially important requirements in more detail.

8. References

- Ambler, S.W. 2002. Agile modeling - Effective Practices for eXtreme Programming and the Unified Process. New York: Wiley Computer Publishing.
- Arora, A., Hall, D., Pinto, C.A., Ramsey, D. and Telang, R. 2004. An ounce of prevention vs. a pound of cure: How can we measure the value of IT security solutions? Carnegie Mellon CyLab.
- Beck, K. 2000. Extreme programming explained. Upper Saddle River: Addison-Wesley.
- Berander, P. 2004. Prioritization of Stakeholder Needs in Software Engineering. Understanding and Evaluation. Licentiate Thesis, Blekinge Institute of Technology, Sweden, Licentiate Series No 2004:12.
- Berander, P., and Jönsson, P. 2006. Hierarchical Cumulative Voting (HCV) - Prioritization of Requirements in Hierarchies. *Int. J. of Software Eng. and Knowledge Eng.* 16(6): 819-849.
- Carver, J., Shull, F., and Basili, V. 2003. Observational Studies to Accelerate Process Experience in Classroom Studies: An Evaluation. *Proc. of the 2003 Int. Symposium on Empirical Software Eng. ISESE*. Rome, Italy, 72-79.
- Cook, T.D., and Campbell, D.T. 1979. Quasi-Experimentation – Design and Analysis Issues for Field Settings. Boston: Houghton Mifflin Company.

- Davis, A.M. 2003. The Art of Requirements Triage. IEEE Computer 36(3): 42 - 49.
- Denne, M., and Cleland-Huang, J. 2003. Software by Numbers: Low-Risk, High-Return Development. Prentice-Hall.
- Dutoit, A.H., and Paech, B. 2002. Rationale-based use case specification. Requirements Eng. J. 7: 3-19.
- Dutoit, A.H., and Paech, B. 2003. Eliciting and maintaining knowledge for requirements evolution. Aurum, A., Jeffery, R., Wohlin, C., and Handzic M. (eds) Managing Software Engineering Knowledge. Berlin: Springer: 135-156.
- Feather, M.S., Cornford, S.L. 2003. Quantitative risk-based requirements reasoning. Requirements Eng. J. 8(4): 248-265.
- Feather, M.S., Cornford, S.L., Kiper, J.D., and Menzies, T. 2006. Experiences using Visualization Techniques to Present Requirements, Risks to Them, and Options for Risk Mitigation. Proc. Int. Workshop on Requirements Eng. Visualization, Minneapolis/ St. Paul, Minnesota.
- Herrmann, A., and Paech, B. 2005. Quality Misuse. Proc. 11th Int. Workshop on Requirements Eng. for Software Quality, Foundations of Software Quality REFSQ, Essener Informatik Beiträge, Band 10: 193-199.
- Herrmann, A., and Paech, B. 2006. Benefit Estimation of Requirements Based on a Utility Function. Proc. 12th Int. Workshop on Requirements Eng. for Software Quality, Foundations of Software Quality REFSQ, Essener Informatik Beiträge, Band 11: 249-250.
- (Herrmann, 2006a) Herrmann, A., Rückert, J., and Paech, B. 2006. Exploring the Interoperability of Web Services using MOQARE. Proc. IS-TSPQ First Int. Workshop on Interoperability Solutions to Trust, Security, Policies and QoS for Enhanced Enterprise Systems. Bordeaux, France.
- Herrmann, A., and Paech, B. 2007. MOQARE: Misuse-oriented Quality Requirements Engineering. to be published by RE Journal.
- Höst, M., Regnell, B., and Wohlin, C. 2000. Using Students as Subjects - A Comparative Study of Students and Professionals in Lead-Time Impact Assessment. Empirical Software Eng. 5(3): 201-214.
- International Standards Organization. 2002. ISO, Risk management – Vocabulary – Guidelines for use in standards, ISO Guide 73. Geneva: International Standards Organization.
- Karlsson, J. 1996. Software requirements prioritising. Proc. 2nd Int. Conf. Requirements Eng., 110-116.
- Karlsson, J., Wohlin, C., and Regnell, B. 1998. An evaluation of methods for prioritizing software requirements. Information and Software Technology 39: 939-947.
- Karlsson, L., Berander, P., Regnell, B., and Wohlin, C. 2004. Requirements Prioritisation: An Experiment on Exhaustive Pair-Wise Comparisons versus Planning Game Partitioning. Berander, P. Prioritization of Stakeholder Needs in Software Engineering, Understanding and Evaluation, Doctoral Thesis, Blekinge Institute of Technology, Licentiate Series No 2004:12.
- Karlsson, L., Thelin, T., Regnell, B., Berander, P., and Wohlin, C. 2007. Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques. Empirical Software Eng. 12(1): 3-33.
- Kontio, J. 1996. The Riskit Method for Software Risk Management, version 1.00. University of Maryland. College park, MD, Computer Science Technical Reports CS-TR-3782.
- Leffingwell, D., and Widrig, D. 2000. Managing Software Requirements - A Unified Approach. Reading, Massachusetts, USA: Addison-Wesley.
- Mayer, N., Rifaut, A., and Dubois, E. 2005. Towards a Risk-Based Security Requirements Engineering Framework. Proc. 11th Int. Workshop on Requirements Eng. for Software

- Quality, Foundations of Software Quality REFSQ, Essener Informatik Beiträge, Band 10, 89-104.
- Ngo-The, A., and Ruhe, G. 2005. Decision Support in Requirements Engineering. Aurum, A., and Wohlin, C. (Eds.) Engineering and Managing Software Requirements. Berlin, Heidelberg: Springer.
- Papadacci, E., Salinesi, C., and Rolland, C. 2004. Payoff Analysis in Goal-Oriented Requirements Engineering. Proc. 10th Int. Workshop on Requirements Eng. for Software Quality, Foundations of Software Quality REFSQ.
- Park, J., Port, D., Boehm, B., and In, H. 1999. Supporting Distributed Collaborative Prioritization for WinWin Requirements Capture and Negotiations. Proc. Int. 3rd World Multiconference on Systemics, Cybernetics and Informatics SCI'99, Vol.2: 578-584.
- Raiffa, H., Richardson, J., and Metcalfe, D. 2002. Negotiation analysis - the science and art of collaborative decision making. Cambridge: Belknap.
- Regnell, B., Höst, M., Natt och Dag, J., Beremark, P., and Hjelm, T. 2001. An Industrial Case Study on Distributed Prioritisation in Market-Driven Requirements Engineering for Packaged Software. Requirements Eng. 6: 51–62.
- Remus, W. 1989. Using Students as Subjects in Experiments on Decision Support Systems. Proc. 22nd Annual Hawaii Int. Conf. on System Sciences, Vol. III: Decision Support and Knowledge Based Systems Track, 176-180.
- Richardson, R. 2003. 2003 CSI/FBI Computer Crime and Security Survey. Computer Security Institute. http://i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2003.pdf (last visit: nov 07)
- Robson, C. 2002. Real World Research. Cornwall, UK: Blackwell Publishing.
- Ruhe, G., Eberlein, A., and Pfahl, D. 2003. Trade-Off Analysis For Requirements Selection. Int. J. Software Eng. And Knowledge Eng. 13(4): 345-366.
- Runeson, P. 2003. Using students as experiment subjects—an analysis on graduate and freshmen student data. Proc. Int. Conf. Empirical Assessment and Evaluation in Software Eng. EASE Keele, UK, 95–102.
- Ryan, K., and Karlsson, J. 1997. Prioritizing Software Requirements in an Industrial Setting. Proc. Int. Conf. on Software Eng., 564-565.
- Saaty, T.L. 1980. The Analytic Hierarchy Process. New York: McGraw-Hill.
- Sindre, G., and Opdahl, A.L. 2000 Eliciting Security Requirements by Misuse Cases. Proc. TOOLS Pacific 2000: 120-131.
- Sindre, G., and Opdahl, A.L. 2001. Templates for Misuse Case Description. Proc. 7th Int. Workshop on Requirements Eng.: Foundation of Software Quality – REFSQ, Essener Informatik Beiträge Band 6. Essen, Germany, 125-136.
- Stamatis, D.H. 2003. Failure Mode and Effect Analysis - FMEA from Theory to Execution. Milwauki, USA: American Society for Quality Press.
- Stylianou, A.C., Kumar, R.L., and Khouja, M.J. 1997. A total quality management-based systems development process. ACM SIGMIS Database 28(3): 59-71.
- Sisyphus <http://sysiphus.in.tum.de/> (last visit: nov 2007)
- Tversky, A., Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. Science 185: 1124–1131.
- Wolf, T., and Dutoit, A.H. 2004. A rationale-based analysis tool. Proc. 13th Int. Conf. on Intelligent on Adaptive Systems and Software Eng.. Nice, France
- Xie, N., Mead, N. R., Chen, P., Dean, M., Lopez, L., Ojoko-Adams, D., and Osman, H. 2004. SQUARE Project: Cost/Benefit Analysis Framework for Information Security Improvement Projects in Small Companies. Software Engineering Institute, Carnegie Mellon University, Technical Note CMU/SEI-2004-TN-045

Annex A: Experiment Material

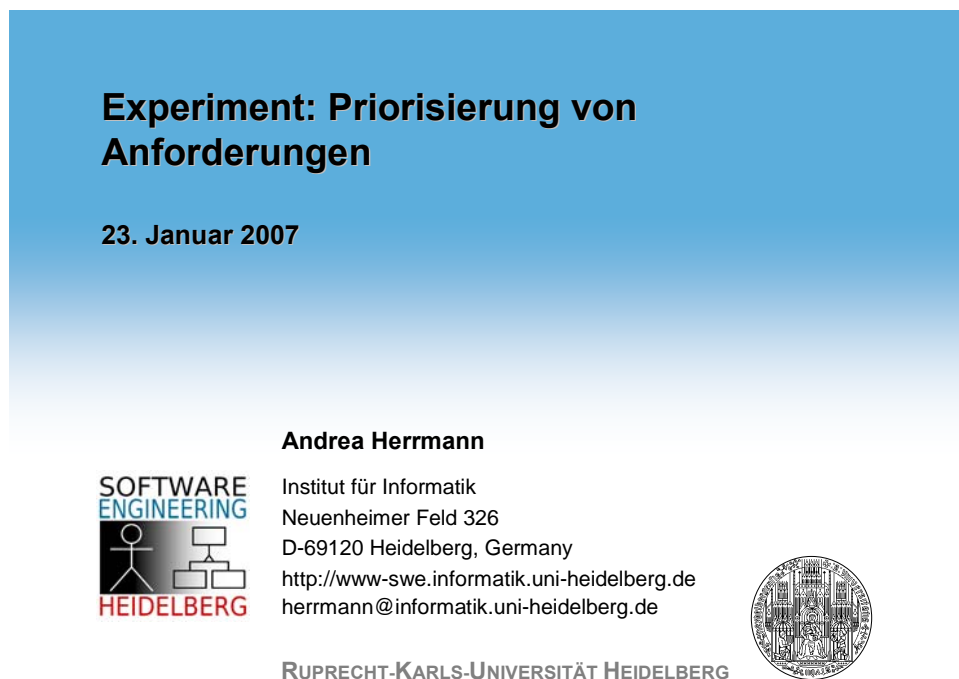
This annex contains the original experiment material which was used during the two experiments respectively. This material includes:

- The presentation slides which were used to present the method and the experiment to the participants
- The questionnaires (tbd: mit misuse cases und misuse tree?)
- The spreadsheets tables used for the experiment results

This material originally is completely in German. Upon request, it can be translated into English for later versions of this technical report.

A.1 Experiment 1

a) Presentation slides





Experiment: Priorisierung von Anforderungen

23. Januar 2007

Andrea Herrmann
Institut für Informatik
Neuenheimer Feld 326
D-69120 Heidelberg, Germany
<http://www-swe.informatik.uni-heidelberg.de>
herrmann@informatik.uni-heidelberg.de

SOFTWARE ENGINEERING HEIDELBERG



RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG

Ziel des Experiments

Experiment

- ▶ Ziel
- Zeitplan
- Fallstudie
- MOQARE

Priorisierung von Anforderungen in MOQARE:

- Wie gut?
- Wie?

Methoden:

1. Intuitive Sortierung
2. MOQARE
3. MOQARE

Slide 2 explains the experiment's goal : We wanted to find out how well requirements can be prioritized in MOQARE and how this is best done. Therefore, three different methods are tested: intuitive ranking and two different methods based on MOQARE. (The students know the method MOQARE already from the former part of the lecture.)

Methode 1: intuitive Sortierung

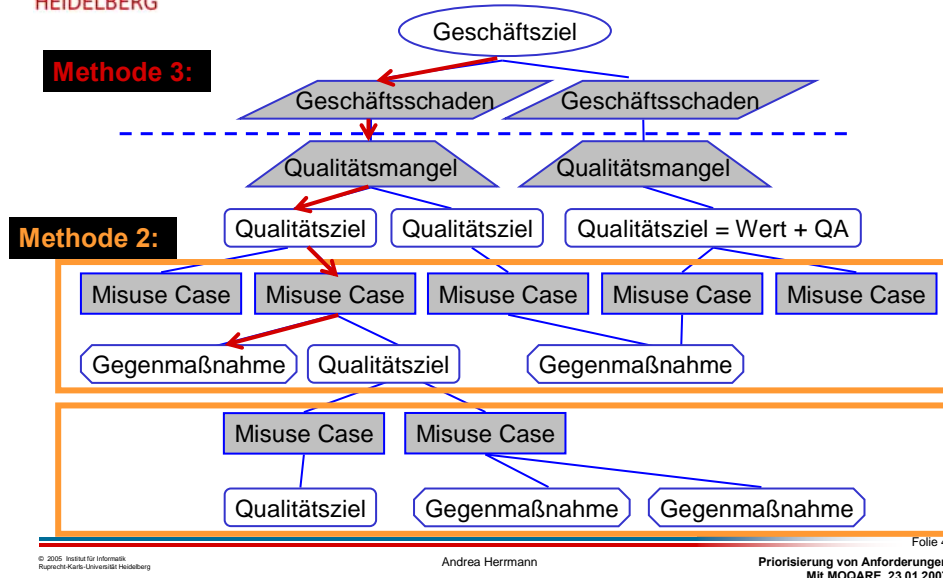
Experiment

- ▶ Ziel
- Zeitplan
- Fallstudie
- MOQARE

1. Intuitive Sortierung in zwei Schritten

- Grobsortierung
- Feinsortierung

Explanation of method 1.



Explanation of methods 2 and 3: Method 2 estimated a countermeasure's benefit relative to the misuse cases, while method 3 starts from the business goals.

	08:30-08:50	Einführung
Experiment	08:50-09:00	Aufgabe 1: 1000€ Methode
Ziel	09:00-09:10	Wh: Priorisierung in MOQARE
Zeitplan	09:10-09:50	Aufgabe 2: Risiken der Misuse Cases -> Prioritäten der Gegenmaßnahmen
Fallstudie	09:50-09:55	Pause
MOQARE	09:55-10:35	Aufgabe 3: Geschäftsziel -> ... -> Prioritäten der Gegenmaßnahmen
	10:35-10:45	Aufgabe 4: Fragen zum Experiment
	Folgetage	Auswertung der Ergebnisse
	30.01.	Diskussion der Ergebnisse

© 2005 Institut für Informatik, Ruprecht-Karls-Universität Heidelberg
Andrea Herrmann
Priorisierung von Anforderungen Mit MOQARE, 23.01.2007
Folie 5

Time table of the experiment. Aufgabe 1 (task 1) is the execution of method 1, Aufgabe 2 of method 2 and Aufgabe 3 of method 3. Aufgabe 4 means the questionnaire 4 with questions concerning the experiment. During the following days, the results are evaluated, and one week later, the post-test took place with discussion of the results.

Online-Flohmarkt, siehe Vorlesung SWE IIc, Teil 3
Funktionale Anforderungen (FR):

- Experiment
- Ziel
 - Zeitplan
 - -Fallstudie
 - MOQARE

Verkäufer = Anbieter

- Waren anbieten
- Verkauf
- liefern

Käufer

- Handeln
- Kauf
- Bezahlung

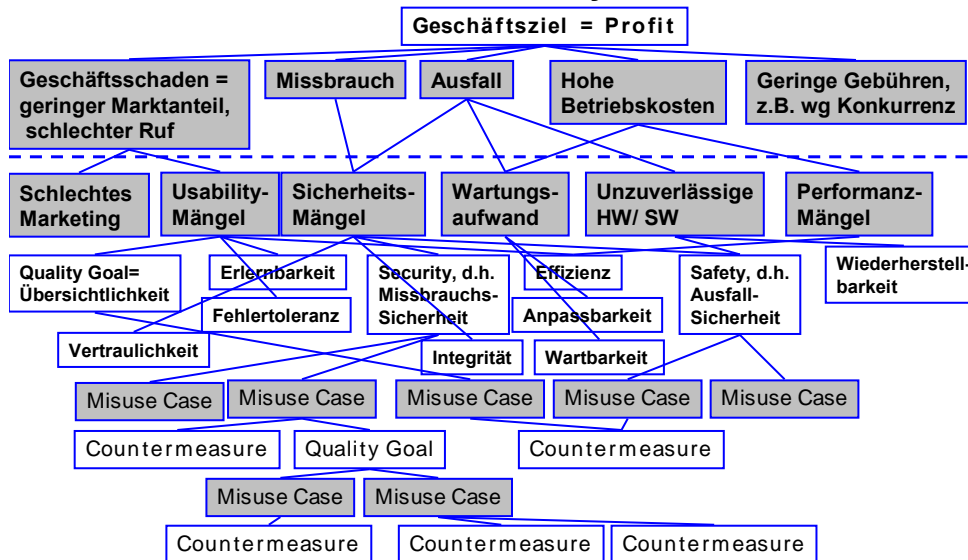
Benutzer = Käufer und Verkäufer

- Registrierung

This slide refers to the flea market example already treated in the lecture in unit 3, and here the functional requirements are recapitulated, described by the actors (=user roles) and the tasks which they execute.

Fallstudie: s Vorlesung SWE IIc, Teil 3

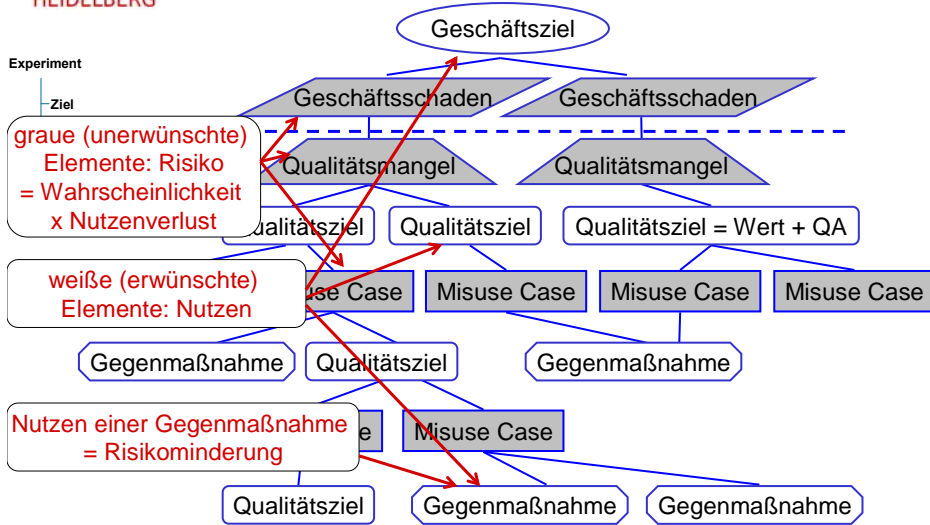
Online-Flohmarkt -> Quality Goals



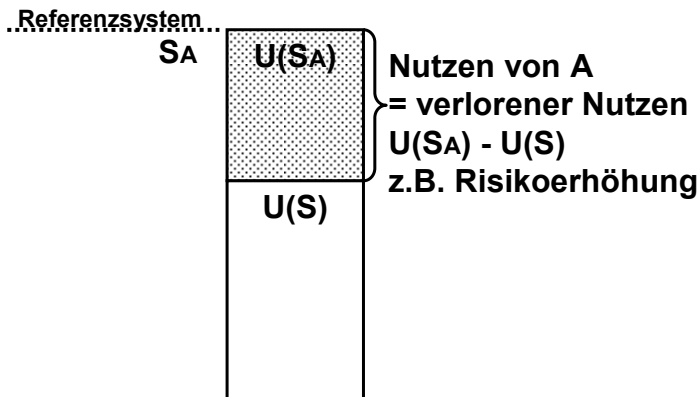
This MOQARE misuse tree also was already known to the experiment participants because it is the sample solution of a home work and we have discussed it before. Therefore, this slide signifies only a recapitulation.

- Experiment
- Ziel
 - Zeitplan
 - Fallstudie
 - ▶ MOQARE

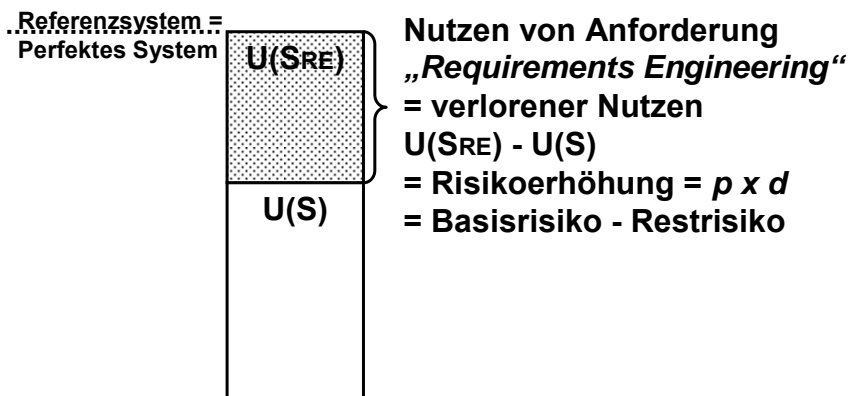
This slide has the function to announce a new chapter.



Some introductory words concerning prioritization in MOQARE: In MOQARE, we use the Misuse Tree as basis for the requirements prioritization. The grey concepts are the unwanted elements, like the Misuse Cases. Their importance is measured by their risk, which is defined as *probability times benefit loss*. The white, wanted elements are prioritized by their benefit. The benefit which a countermeasure adds is that it reduces Misuse Case risk. Therefore, its benefit is estimated based on risk reduction relative to the Misuse Cases.

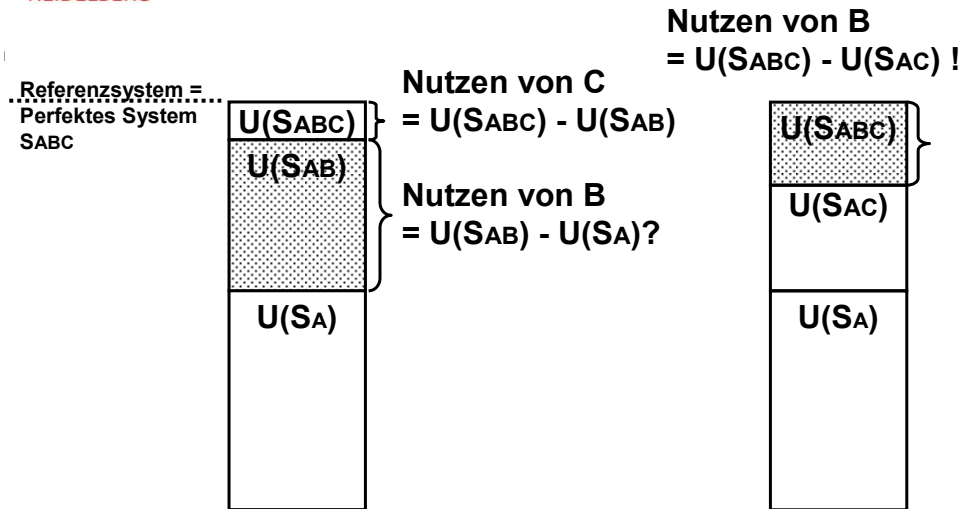


This slide explains the principle of the reference system: Benefit of a single countermeasure A is measured by the benefit lost = risk increase when it is not implemented with respect to the reference system. (As the reference system is the perfect system, we need to consider only this case here.)



Here, an example is treated, which will not appear in the experiment. If “requirements engineering” is a countermeasure, then the lost benefit is *basis risk minus residual risk*.

Referenzsystem: Achtung!



Nutzen von A + Nutzen von B + Nutzen von C \neq $U(S_{ABC})$!

Folie 12
rungen
01.2007

Here, the importance of using the reference system is emphasized. The benefit of B is not estimated by deducing it from System S_{AB} , but from S_{ABC} . This is usually not equivalent, as is shown graphically here.

Vielen Dank für eure Aufmerksamkeit!

Experiment

- Ziel
- Zeitplan
- Fallstudie
- MOQARE

Andrea Herrmann, Barbara Paech

Institut für Informatik
Neuenheimer Feld 326
D-69120 Heidelberg
Germany

<http://www-swe.informatik.uni-heidelberg.de>
{herrmann, paech}@informatik.uni-heidelberg.de

b) Material: Handouts and Questionnaires

As was described in chapter 4.1 Experiment No. 1: Preparation, the experiment participants received the following six handouts and questionnaires:

- the introduction
- Questionnaire Q1 supports method 1
- Questionnaire Q2 supports method 2.
- Questionnaire Q3 supports method 3.
- Questionnaire Q4, distributed directly after the experiment
- Questionnaire Q5, distributed one week after the experiment, during the post-test session

In the subsequent sub-chapters, these six documents will be presented, after a short explanation for the readers of this report, which are printed in *italic*.

A.b.1 Introduction

Explanation for the reader of this report: The introduction handout describes the goal of the experiment and the case study, including information about competitors, the company, project execution and staff, and the flea market's functional requirements.

Fallstudienbeschreibung

Ziel des Experiments:

Momentan entwickelt der Lehrstuhl für Software Engineering eine Methode zur Priorisierung von nicht-funktionalen Anforderungen (NFR). NFR werden bei uns mit MOQARE hergeleitet und dokumentiert. Es sind für diese Form der Anforderungen unterschiedlich aufwändige Möglichkeiten der Priorisierung denkbar. In dem hier vorgeschlagenen Experiment sollen zwei verschiedene Varianten auf ein Beispiel angewendet und deren Ergebnisse miteinander verglichen werden. In Aufgabe 2) wird der Nutzen der Gegenmaßnahmen aufgrund der Risiko-Bewertung der Misuse Cases ermittelt, in Aufgabe 3) werden alle Bewertungen aus den Geschäftszielen hergeleitet. Zum Vergleich wird zunächst in Aufgabe 1) eine intuitive Priorisierungsmethode eingesetzt, bei der Anforderungen zunächst grob und dann fein sortiert werden.

Aufgabenbeschreibung:

Es sollen im Folgenden mit den oben genannten drei Methoden einige Anforderungen einer Fallstudie priorisiert werden. Das Ziel dieser Priorisierung soll es sein, die Lösung von Anforderungskonflikten vorzubereiten. Können während des Projektes aus technischen, finanziellen oder organisatorischen Gründen nicht alle Anforderungen in einer ersten Produktversion gleich gut realisiert werden, sollen die Prioritäten die Entscheidung unterstützen, welche der Anforderungen besonders gut erfüllt werden sollen. Es ist davon auszugehen, dass auch niedrig priorisierte Anforderungen zu einem gewissen geringen Grad erfüllt werden, so weit es ohne speziellen Mehraufwand möglich ist.

Umfeld: Eine Firma plant, einen neuen Internet-Flohmarkt zu entwickeln und anzubieten, auf dem Privatpersonen neue und gebrauchte Waren an andere Privatpersonen verkaufen können. Dieser soll Konkurrenz sein zu beispielsweise www.ebay.de, www.hood.de, www.amazon.de, www.markt.de, www.kleinanzeigen-landesweit.de, www.lass-es-mir.de/, www.quoka.de/, www.zum-flohmarkt.de und so weiter. Um sich von der Konkurrenz abzuheben, soll besonders die Ähnlichkeit mit einem echten Flohmarkt größer sein als allgemein üblich.

Das Projekt wird von einer Firma durchgeführt, die bisher von Web Hosting und Webseitenerstellung lebt. Diese Firma hat 20 Mitarbeiter/innen und betreibt 4 Unix-Server. Der Flohmarkt soll neben den normalen Tätigkeiten her realisiert werden.

Die funktionalen Anforderungen an den Internet-Flohmarkt, wie wir sie früher in der Übung und in Hausaufgaben ermittelt haben, liegen bei (siehe Anlage). Einige nicht-funktionale Anforderungen bzw. Gegenmaßnahmen sollen in diesem Experiment mit drei verschiedenen Methoden priorisiert werden. Wichtig ist hierbei, die Aufgaben 1 bis 4 in der vorgesehenen Reihenfolge durchzuführen.

Diese neun **Anforderungen/ Gegenmaßnahmen** sollen priorisiert werden:

- **A1:** Benutzeroberfläche übersichtlich und intuitiv gestalten, z.B. durch aussagekräftige Beschriftungen, Befolgen von Usability-Richtlinien
- **A2:** Support – den Benutzern Support über verschiedene Medien anbieten (per Telefon und E-Mail)
- **A3:** Ähnlichkeit mit einem echten Flohmarkt
- **A4:** Inspektion der Spezifikationsdokumente
- **A5:** Verschlüsselte Speicherung der Kundendaten
- **A6:** Schnelle Hard- und Software
- **A7:** Standards – Einhalten von Standards bei der Programmierung sowie dem Design der Benutzeroberfläche
- **A8:** Automatisierte Benachrichtigung der Service-Mitarbeiter bei Ausfall des Systems
- **A9:** Ersatz-Server

Anlage:

- Funktionale Anforderungen

Anlage:

Funktionale Anforderungen an den Online-Flohmarkt

(siehe Musterlösung von Hausaufgabe 3.2)

Rollen, Tasks und Use Cases:

Verkäufer = Anbieter

- Waren anbieten
 - Suche nach Inhalten des Online Marktes
 - Eingabe von Warendaten
 - Aktualisierung von Warendaten
 - Angebot herausnehmen
- Verkauf
 - Zahlungseingangsprüfung und Versand
 - Kontakt mit Käufer aufnehmen
 - Beschwerden bearbeiten
 - Käufer bewerten
- Liefern
 - Zahlungseingang prüfen
 - Ware versenden

Käufer

- Handeln
 - Suche nach Inhalten des Online Marktes
 - Stöbern
 - Verhandeln über den Preis
- Kauf
 - Weitere Information über Angebote einholen
 - Ware kaufen
 - Zahlungsweise und Lieferadresse angeben
 - Verkaufsstatus beobachten
 - Mit Verkäufer Kontakt aufnehmen
 - Verkäufer bewerten
- Bezahlung
 - Ware bezahlen

Benutzer = Käufer und Verkäufer

- Registrierung
 - Eingabe der Registrierungsdaten
 - Persönliche Daten pflegen
 - Eventuell Post Ident Verfahren

Besucher

- Besuch
 - Suche nach bestimmten Waren
 - Stöbern

System Administrator

- Systemadministration
 - Zugriffsrechte Verwaltung
 - Backuperstellung
 - Datenwiederherstellung
 - Beobachten und Sicherstellen der Verfügbarkeit des Systems
 - Korrektur von Softwarefehlern
 - Einspielen von Updates
 - Benutzer sperren
 - Benutzer löschen

Geschäftsführer

- Controlling des Profits
 - Vergleich des tatsächlichen mit dem realen Profit

Marketing, Vertrieb, etc.

- Online Werbung
 - Benutzungsprofile erstellen
 - Benutzungsprofile auswerten
 - Etc.

Redakteur

- Datenpflege
 - Datenaktualisierung

Entwickler

- Erweiterung und Optimierung des Systems

Bank

- Überweisung buchen

Lieferdienst

- Ware zustellen

A.b.2 Questionnaire Q1

Explanation for the reader of this report: Questionnaire Q1 supports method 1 by a table which allows to attribute a group and a priority to each requirement. It also asks how certain the participant feel about their judgements.

Name: _____

Allgemeine Fragen

Kreuze das Zutreffende an:

- Du warst dabei, als in der Vorlesung gemeinsam die Qualitätsziele des Online-Flohmarkts hergeleitet wurden (in der Vorlesung am 31.10.2006).
- Du hast Hausaufgabe 3.1 zur Ermittlung der nicht-funktionalen Anforderungen (NFR) durchgeführt.
- Du hast Hausaufgabe 3.2 zur Ermittlung der funktionalen Anforderungen (FR) durchgeführt.
- Du warst dabei, als die Ergebnisse der Hausaufgabe besprochen wurden (Beginn der Vorlesung am 07.11.2006).

Aufgabe 1: Intuitive Sortierung in zwei Schritten

Priorisiere die in der Einführung bereits genannten neun Anforderungen, indem du im ersten Schritt jede einordnest in die Gruppen: „hoher Nutzen“, „durchschnittlich nützlich“ oder „geringer Nutzen“.

Im zweiten Schritt beurteilst du innerhalb jeder der Gruppen die einzelnen Anforderungen, so dass du ein Gesamt-Ranking erhältst. Wähle die Zahl 1 für die wichtigste und 9 für die unwichtigste Anforderung.

Startzeit: _____

Anforderung	Grobbewertung: "hoher Nutzen", "durchschnittlich nützlich", "geringer Nutzen"	Feinbewertung
A1: Benutzeroberfläche übersichtlich und intuitiv gestalten, z.B. durch aussagekräftige Beschriftungen, Befolgen von Usability-Richtlinien		
A2: Support		
A3: Ähnlichkeit mit einem echten Flohmarkt		
A4: Inspektion der Spezifikationsdokumente		
A5: Verschlüsselte Speicherung der Kundendaten		
A6: Schnelle Hard- und Software		
A7: Standard		
A8: Automatisierte Benachrichtigung der Service-Mitarbeiter bei Ausfall des Systems		
A9: Ersatz-Server		

Tabelle 1.1

Endzeit: _____

Wie sicher warst du dir bei deinen Beurteilungen?

- Sehr sicher
- Eher sicher
- Gemischt
- Eher unsicher
- Sehr unsicher

Gab es Anforderungen, bei denen du besonders unsicher bist, sie richtig eingeordnet zu haben? Wenn ja: Welche? Und welche Informationen haben dir eventuell gefehlt?

Anhand welcher Kriterien hast du die Anforderungen explizit oder implizit bewertet?

A.b.3 Questionnaire Q2

Explanation for the reader of this report: Questionnaire Q2 supports method 2. It first informs about the expected revenues and costs of the system and that all estimations are to be done for a period of two years. The reference system is defined to be the perfect system where all countermeasures are supposed to be implemented. Method 2 is supported by two tables: one for the reference risk and the other for the varied risk, each containing a column for the probability and for the relative damage estimation. Neither misuse case risk nor countermeasure benefit are calculated here. The participants are asked what relative uncertainty (in % points) they expect for their probabilities and damages and whether there were misuse cases where they were especially uncertain (variables e and f).

Name: _____

Aufgabe 2: Priorisierung von Gegenmaßnahmen, ausgehend von Risiken der Misuse Cases

Das Vorgehen in Aufgabe 2 ist folgendes: Es werden dieselben Anforderungen (Gegenmaßnahmen) des Online-Flohmarkts wie in Aufgabe 1 priorisiert, jetzt unter Verwendung der Beziehungen zwischen Gegenmaßnahmen und Misuse Cases. Gehe hierbei Schritt für Schritt entsprechend der folgenden Anweisungen vor. (Die konkreten Arbeitsanweisungen sind kursiv.)

Um eine Richtgröße für die Abschätzung der entstehenden Schäden zu haben, erhältst du folgende Informationen: Die Firma plant Einnahmen durch Einstellgebühren (pro Woche 20.000 x 10 Cent), aber auch durch Anzeigen von Werbekunden (pro Woche 35 Anzeigen à 100€), d.h. zusammen Einnahmen von 5.500€ pro Woche. Berücksichtigt man Betriebskosten von 500€ pro Woche (der niedrige Betrag begründet sich aus Synergieeffekten mit den anderen Tätigkeiten der Firma), kann man mit einem Gewinn von 5000€ pro Woche rechnen, d.h. rund 250.000€ pro Jahr.

Bisher macht diese Firma einen Umsatz von 1,5 Millionen € pro Jahr, davon 100.000 € Gewinn. Die Kosten für die Erstellung des Internet-Flohmarkts betragen schätzungsweise eine halbe Million. Diese kann das System erwartungsgemäß in zwei Jahren einspielen. Alle Kosten und Risiken sollen im Folgenden daher ebenfalls auf eine **Dauer von zwei Jahren** bezogen berechnet werden.

Gäbe es keinen Umsatz, wäre der **Gesamtschaden** 550.000€ (=500.000€ für die Erstellung der Software plus die Betriebskosten für zwei Jahre, d.h. rund 100 Wochen mal 500€).

Das Referenzsystem sei das perfekte System, in dem alle funktionalen Anforderungen und die Gegenmaßnahmen A1-A9 erfüllt sind. Trotz aller Gegenmaßnahmen bleibt für viele Misuse Cases noch ein **Restrisiko** bestehen. Ermittle zunächst das Restrisiko für die Misuse Cases im Referenzsystem. *Schätze hierzu für jeden Misuse Case die Wahrscheinlichkeit p in % sowie den erwarteten Schaden d anteilig in % vom Gesamtschaden von 550.000€.* Da wir hier nur mit relativen Werten rechnen, kannst du die Wahrscheinlichkeiten und Schäden auf einen Verkauf-/ Kauf-Vorgang beziehen, beispielsweise bei MUC 1: Wie hoch ist die Wahrscheinlichkeit, dass bei einem versuchten Kauf dieser MUC passiert und welchen Anteil des Profits durch diesen Verkauf verliert man dadurch? Misuse Case 7 dagegen bezieht sich beispielsweise nicht auf einen Verkaufsvorgang, sondern auf den kontinuierlichen Betrieb.

Wie hoch ist die Wahrscheinlichkeit für diesen Misuse Case während der Betriebszeit von zwei Jahren?

Das Risiko ist definiert als das Produkt aus Wahrscheinlichkeit und Schaden $p \times d$ und wird von uns später während der Auswertung berechnet.

Startzeit der Restrisikoschätzung: _____

Restrisiko pro Misuse Case im Referenzsystem, in dem alle funktionalen Anforderungen und die Gegenmaßnahmen A1-A9 realisiert sind:

Misuse Case	Wahrscheinlichkeit p des Misuse Case in %	Schaden anteilig am Nutzen des Geschäftsziels (550.000€), in %
MUC1: Benutzerfehler vereitelt geplanten Kauf		
MUC2: Vernachlässigung von Übersichtlichkeitsanforderungen bei der Softwareentwicklung führt zu Kundenverlust		
MUC3: Verkäufer gibt nach längerer Zeit entnervt auf, einen Artikel einstellen zu wollen		
MUC4: Benutzer ohne technischem Hintergrund verstehen technische Begriffe/ Oberfläche nicht - > lange Lernphase und Kundenverlust		
MUC5: Programm hilft Benutzer nicht bei Fehleingaben		
MUC6: Kundendaten werden von nicht autorisierter Person gelesen		
MUC7: Hacker manipulieren die Seite samt ihrer Inhalte		
MUC8: Ineffizienz durch lange Antwortzeiten des Systems		
MUC9: Komplexität des Systems führt zu hohem Wartungsaufwand		
MUC10: geringe Benutzereffizienz durch schlechtes Auffinden der Information, Nichterkennen von Wesentlichem		
MUC11: Code kann bei Änderungen im Systemumfeld nicht oder nur aufwändig wiederverwendet werden		
MUC12: Der einzige Server fällt aus, erst nach Stunden bemerkt ein Service-Mitarbeiter den Ausfall durch Zufall; ein Ersatz-Server ist nicht vorhanden; die Benutzer können tagelang nicht auf die Seite		

Tabelle 2.2

Wie unsicher warst du dir jeweils bei der Bewertung der Wahrscheinlichkeiten? (Liegt ein Schätzwert p von 10% vermutlich zwischen 5% und 15%, sind das ± 5 Prozentpunkte.) Wenn du bei einzelnen Werten besonders unsicher oder sicher warst, nenne diese extra.

Wie unsicher warst du dir jeweils bei der Bewertung der entstehenden Schäden? (Liegt ein Schätzwert von 10% vermutlich zwischen 5% und 15%, sind das ± 5 Prozentpunkte.)
Wenn du bei einzelnen Werten besonders unsicher oder sicher warst, nenne diese extra.

Endezeit der Restrisikoschätzung: _____

Um den Nutzen der Gegenmaßnahmen zu ermitteln, wird nun das **Basisrisiko** abgeschätzt. Das Basisrisiko misst das Risiko ohne Gegenmaßnahme. Hier interessieren uns Risiken für Systeme, die sich vom Referenzsystem dadurch unterscheiden, dass eine einzelne Gegenmaßnahme nicht realisiert ist. Der Nutzen einer Gegenmaßnahme berechnet sich später daraus, um wie viel sich durch ihre Nicht-Realisierung ein oder mehrere Risiken erhöhen.

Um jeweils das Basisrisiko abzuschätzen, gehe in der folgenden Tabelle zeilenweise vor. Stelle dir vor, im Gegensatz zum Referenzsystem sei die eine angegebene Gegenmaßnahme nicht implementiert. Wie hoch sind dann jeweils Wahrscheinlichkeit und Schaden des angegebenen Misuse Case? Gib den Schaden im prozentualen Verhältnis zu dem (Rest-) Schaden in Tabelle 2.1 an, d.h. wenn der Schaden anderthalb mal so hoch ist, trage 150 ein. Das Risiko berechnen wir später während der Auswertung.

Startzeit der Basisrisikoschätzung: _____

Basisrisiko pro Misuse Case in einem System, in dem alle funktionalen Anforderungen und die Gegenmaßnahmen A1-A9 realisiert sind, bis auf eine:

Misuse Case	Nicht realisierte Gegenmaßnahme	Wahrscheinlichkeit p des Misuse Case in %	Schaden relativ zum Restschaden in Tab. 2.1, in %
MUC1: Benutzerfehler vereitelt geplanten Kauf	A2		
MUC1: Benutzerfehler vereitelt geplanten Kauf	A1		
MUC2: Vernachlässigung von Übersichtlichkeitsanforderungen bei der Softwareentwicklung führt zu Kundenverlust	A1		
MUC3: Verkäufer gibt nach längerer Zeit entnervt auf, einen Artikel einstellen zu wollen	A2		
MUC4: Benutzer ohne technischem Hintergrund verstehen technische Begriffe/ Oberfläche nicht -> lange Lernphase und	A3		
MUC5: Programm hilft Benutzer nicht bei Fehleingaben	A4		
MUC6: Kundendaten werden von nicht autorisierter Person gelesen	A5		
MUC7: Hacker manipulieren die Seite samt ihrer Inhalte	A5		
MUC8: Ineffizienz durch lange Antwortzeiten des Systems	A6		
MUC9: Komplexität des Systems führt zu hohem Wartungsaufwand	A7		
MUC10: geringe Benutzereffizienz durch schlechtes Auffinden der Information, Nichterkennen von Wesentlichem	A7		
MUC11: Code kann bei Änderungen im Systemumfeld nicht oder nur aufwändig wiederverwendet werden	A7		
MUC12: Der einzige Server fällt aus, erst nach Stunden bemerkt ein Service-Mitarbeiter den Ausfall durch Zufall; ein Ersatz-Server ist nicht vorhanden; die Benutzer können tagelang nicht auf die Seite	A8		
MUC12: Der einzige Server fällt aus, erst nach Stunden bemerkt ein Service-Mitarbeiter den Ausfall durch Zufall; ein Ersatz-Server ist nicht vorhanden; die Benutzer können tagelang nicht auf die Seite	A9		

Tabelle 2.3

Wie unsicher warst du dir jeweils bei der Bewertung der Wahrscheinlichkeiten? (Liegt ein Schätzwert p von 10% vermutlich zwischen 5% und 15%, sind das ± 5 Prozentpunkte.)
Wenn du bei einzelnen Werten besonders unsicher oder sicher warst, nenne diese extra.

Wie unsicher warst du dir jeweils bei der Bewertung der entstehenden Schäden?
Wenn du bei einzelnen Werten besonders unsicher oder sicher warst, nenne diese extra.

Endezeit der Basisrisikoschätzung: _____

Wie bist du bei der Bewertung der Schäden vorgegangen?

Aus deinen obigen Abschätzungen wird während der Auswertung der Nutzen der einzelnen Gegenmaßnahmen berechnet, auf der Grundlage der durch sie verursachten Verringerung der Risiken der Misuse Cases. Die Ergebnisse dieser Berechnung werden wir nächste Woche besprechen.

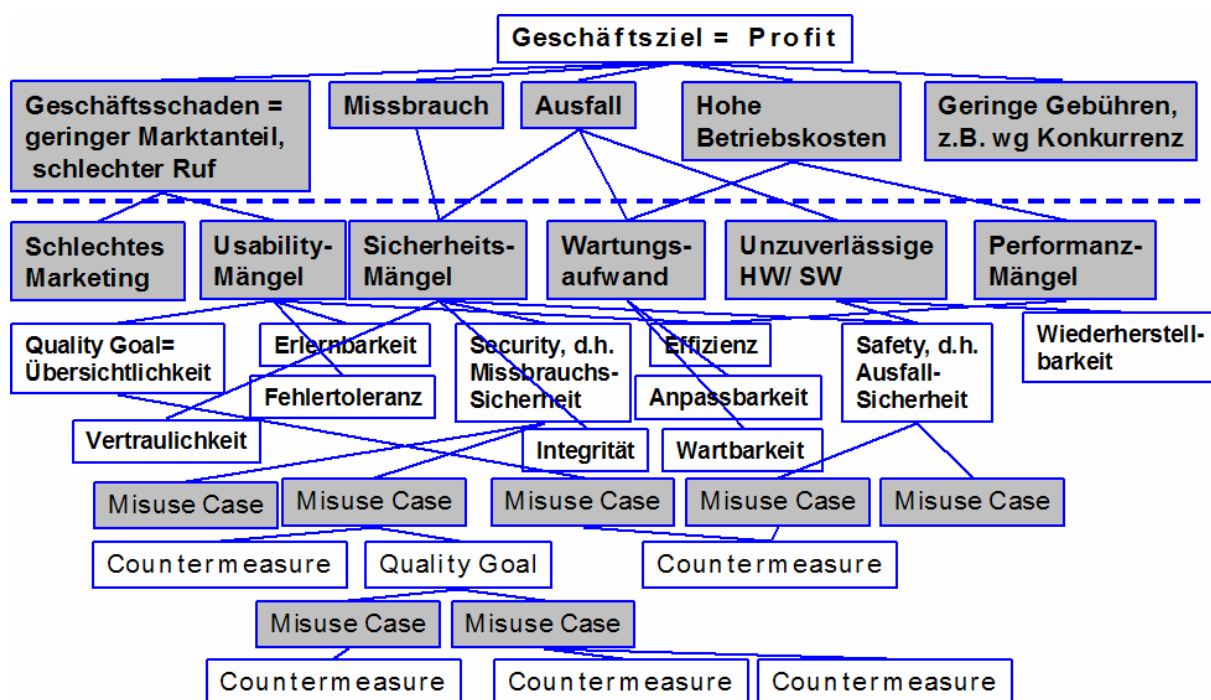
A.b.4 Questionnaire Q3

Explanation for the reader of this report: Questionnaire Q3 supports method 3. Each of the steps described in section 3 is supported by one separate table. Here, exclusively relative values are estimated for probabilities and damages. Estimated values are reused from questionnaire 2 as far as possible. Like in Q2, the participants are asked what relative uncertainty (in % points) they expect for their probabilities and damages and whether there were misuse cases where they were especially uncertain (variables e and f).

Name: _____

Aufgabe 3: Priorisierung von Gegenmaßnahmen, ausgehend von den Geschäftszielen

Leite nun den Nutzen der Gegenmaßnahmen von oben nach unten entlang des Misuse Tree ab, beginnend bei dem Geschäftsziel „Profit“.



Gehe dabei nach den folgenden Schritten vor:

- Setze den Nutzen des **Geschäftsziels** „Profit“ gleich 550.000€. Der Nutzen des Geschäftsziels ist gleich der Antwort auf die Frage: Wie hoch wäre der Schaden, wenn es überhaupt keinen Profit gäbe? In diesem Fallbeispiel sind das die 500.000€ für die

Erstellung der Software plus die Betriebskosten für zwei Jahre (rund 100 Wochen mal 500€).

- Für jeden **Geschäftsschaden** schätze in Tabelle 3.1 ab, zu welchem Anteil (in %) er das Geschäftsziel zerstören würde, wenn er eintritt. Multipliziert mit dem Nutzen des Geschäftsziels ergibt sich der verursachte Schaden. (Diese Berechnung führen wir während der späteren Auswertung durch.) Setze bei den Schätzungen immer das Referenzsystem voraus, d.h. das System, in dem alle funktionalen Anforderungen und die Gegenmaßnahmen A1-A9 realisiert sind. Achtung: Die Summe aller Anteile muss nicht 100% ergeben.
Notiere auch die Unsicherheit deiner Schätzungen.
- Für jeden **Qualitätsmangel** schätze ab, mit welcher Wahrscheinlichkeit er zu dem mit ihm verbundenen Geschäftsschaden führt, wenn er vorhanden ist bzw. wäre.

Beispiel: Wir schätzen, dass der Geschäftsschaden „geringe Gebühren z.B. wegen Konkurrenz“ etwa 30% des Profits kosten würde und sind zu ± 10 Prozentpunkte sicher (d.h. vermutlich kostet dieser Geschäftsschaden 20-40% des Profits). Der zu betrachtende Qualitätsmangel sei „zu grell-bunte Benutzeroberfläche“ und führt mit 70%iger Wahrscheinlichkeit dazu, dass Kunden deswegen zur Konkurrenz abwandern. Dann hat die Tabelle 3.1 die folgenden Einträge:

Geringe Gebühren z.B. wegen Konkurrenz	30	10	zu grell-bunte Benutzeroberfläche	70
--	----	----	--------------------------------------	----

Anfangszeit: _____

Geschäftsschaden	Anteil des Geschäftsziels "Profit", der durch den jeweiligen Geschäfts-schaden zerstört wird, in %	Unsicherheit des Anteils (+/- Prozentpunkte)	Qualitätsmangel	Wahrscheinlichkeit, dass der Qualitäts-mangel zum Geschäfts-schaden führt, in %
Geringer Marktanteil, schlechter Ruf			Usability-Mängel	
Missbrauch des Systems			Sicherheitsmängel	
Ausfall des Systems			Sicherheitsmängel	
			hoher Wartungsaufwand	
			Unzuverlässigkeit der Hardware/ Software	
Hohe Betriebskosten			hoher Wartungsaufwand	
			Performanzmängel	

Tabelle 3.4

Wie unsicher warst Du Dir jeweils bei der Bewertung der Wahrscheinlichkeiten für die Qualitätsmängel? (Liegt ein Schätzwert von 10% vermutlich zwischen 5% und 15%, sind das ± 5 Prozentpunkte.) Wenn du bei einzelnen Werten besonders unsicher oder sicher warst, nenne diese extra.

Endezeit für die Bewertung der Geschäftsschäden und Qualitätsmängel: _____

- Schätze als nächstes den relativen Schaden eines **Qualitätsziels** (in %) am Risiko des Qualitätsmangels:
 - Verursacht ein nicht-erreichtes Qualitätsziel einen zugeordneten Qualitätsmangel vollständig, schreibst du in Tabelle 3.2 „100“.
 - Bei teilweiser Verursachung eines Qualitätsmangels durch ein nicht-erreichtes Qualitätsziel entsprechend anteilig, z.B. 50.

Beispiel:

Wir betrachten drei Qualitätsziele QZ1-QZ3, die zwei Qualitätsmängel QM1-QM2 hervorrufen:

Qualitätsziele	Qualitätsmängel	relativer Schaden des nicht-erreichten Qualitätsziels anteilig am Qualitätsmangel, in %
QZ1	QM1	50
QZ2	QM2	70
QZ3	QM2	60

Wird QZ1 nicht erreicht, sei QM1 halb (=50%) verursacht.

Ist QZ2 nicht erreicht, geschieht QM2 beispielsweise zu 70%, und bei Nicht-Erreichen von QZ3 zu 60%. Die Summen müssen nicht 100 ergeben, da Ursachen und Folgen meist nicht unabhängig sind. Im obigen Beispiel schadet jedes nicht erreichte Qualitätsziel für sich den Qualitätsmangel bereits in hohem Maße; würden beide Qualitätsziele gleichzeitig nicht erreicht, wäre der Qualitätsmangel vielleicht beinahe vollständig erreicht, was hier jedoch gar nicht abgeschätzt werden soll.

Anfangszeit: _____

Qualitätsziel	Qualitätsmängel	relativer Schaden des nicht-erreichten Qualitätsziels am Qualitätsmangel, in %
Übersichtlichkeit der Benutzeroberfläche	Usability-Mängel	
Erlernbarkeit der Benutzer-Tasks	Usability-Mängel	
Fehlertoleranz der Benutzeroberfläche	Usability-Mängel	
Effizienz	Usability-Mängel	
Effizienz	Performanzmängel	
Wartbarkeit	hoher Wartungsaufwand	
Portierbarkeit	hoher Wartungsaufwand	
Safety, d.h. Ausfallsicherheit	Unzuverlässigkeit der HW/SW	
Wiederherstellbarkeit	Unzuverlässigkeit der HW/SW	
Safety, d.h. Ausfallsicherheit	Sicherheitsmängel	
Security, d.h. Missbrauchs-sicherheit	Sicherheitsmängel	
Vertraulichkeit der Kundendaten	Sicherheitsmängel	
Integrität der Daten und Prozesse	Sicherheitsmängel	

Tabelle 3.5 (Anm.: Die Qualitätsziele beziehen sich, wo nicht anders angegeben, auf das gesamte System.)

Für wie unsicher hältst du deine Abschätzungen des Nutzens der Qualitätsziele in Prozentpunkten? Wenn du bei einzelnen Werten besonders unsicher oder sicher warst, nenne diese extra.

Endezeit für die Bewertung der Qualitätsziele: _____

- *Berechne nun für die **Misuse Cases** jeweils ihren zum Restrisiko gehörenden Schaden. Später bei der Auswertung werden wir die in Aufgabe 2) geschätzten Wahrscheinlichkeiten wiederverwenden. Schätze hier den Schaden. Dieser entspricht dem Nutzenverlust, den der Misuse Case am Qualitätsziel verursacht.*

Anfangszeit für die Ermittlung der Restrisiken: _____

Auszufüllen ist in der folgenden Tabelle 3.3 nur die Spalte mit den Schäden. Schätze ab, zu welchem Anteil (in %) ein Misuse Case im Referenzsystem das angegebene Qualitätsziel beschädigt.

Beispiel: Wird das Qualitätsziel „dezent Farben der Benutzer-Oberfläche“ durch den Misuse Case „Entwickler ist farbenblind und verwendet grelle Farben“ verursacht, entsteht vermutlich ein Schaden von (95 ± 5) % am Qualitätsziel. Die Wahrscheinlichkeit wird hier nicht geschätzt, würde sich aber berechnen aus der Wahrscheinlichkeit, dass ein farbenblinder Entwickler für die Entwicklung der Benutzer-Oberfläche eingesetzt wird und dann auch wirklich grelle Farben verwendet. Diese Wahrscheinlichkeit ist normalerweise sehr gering.

Restrisiko pro Misuse Case im Referenzsystem, in dem alle funktionalen Anforderungen und die Gegenmaßnahmen A1-A9 realisiert sind:

(Tabelle 3.6)

Qualitätsziel	Misuse Case	Wahrscheinlichkeit p des Misuse Case in %	Schaden anteilig am Qualitätsziel in %
Übersichtlichkeit der Benutzeroberfläche	MUC2: Vernachlässigung von Übersichtlichkeitsanforderungen bei der Softwareentwicklung führt zu Kundenverlust		
Erlernbarkeit der Benutzer-Tasks	MUC4: Benutzer ohne technischem Hintergrund verstehen Begriffe/ Oberfläche nicht -> lange Lernphase und Kundenverlust		
Fehlertoleranz der Benutzeroberfläche	MUC1: Benutzerfehler vereitelt geplanten Kauf		
Fehlertoleranz der Benutzeroberfläche	MUC5: Programm hilft Benutzer nicht bei Fehleingaben		
Effizienz	MUC3: Verkäufer gibt nach längerer Zeit entnervt auf, einen Artikel einstellen zu wollen		
Effizienz	MUC8: Ineffizienz durch lange Antwortzeiten des Systems		
Effizienz	MUC10: geringe Benutzereffizienz durch schlechtes Auffinden der Information, Nichterkennen von Wesentlichem		
Portierbarkeit	MUC11: Code kann bei Änderungen im Systemumfeld nicht oder nur aufwändig wiederverwendet werden		
Safety = Ausfallsicherheit	MUC12: Der einzige Server fällt aus, erst nach Stunden bemerkt ein Service-Mitarbeiter den Ausfall durch Zufall; ein Ersatz-Server ist nicht vorhanden; die Benutzer können tagelang nicht auf die Seite		
Wartbarkeit	MUC9: Komplexität des Systems führt zu hohem Wartungsaufwand		
Vertraulichkeit der Kundendaten	MUC6: Kundendaten werden von nicht autorisierter Person gelesen		
Integrität der Daten und Prozesse	MUC5: Programm hilft Benutzer nicht bei Fehleingaben		
Integrität der Daten und Prozesse	MUC7: Hacker manipulieren die Seite samt ihrer Inhalte		

Wie unsicher warst du dir jeweils bei der Bewertung der entstehender Schäden? (Liegt ein Schätzwert von 10% vermutlich zwischen 5% und 15%, sind das ± 5 Prozentpunkte.) Wenn du bei einzelnen Werten besonders unsicher oder sicher warst, nenne diese extra.

Endezeit für die Bewertung der Restrisiken: _____

Die **Basisrisiken** brauchst du nicht noch mal abzuschätzen, die lassen sich aus deinen bisherigen Ergebnissen in Aufgabe 2 und 3 herleiten.

- Der Nutzen jeder **Gegenmaßnahme** berechnet sich nun aus der Risikoverminderung. Diese Berechnung gehört zur Auswertung, die wir im Anschluss durchführen und deren Ergebnis wir nächste Woche besprechen.

A.b.5 Questionnaire Q4

Explanation for the reader of this report: Questionnaire Q4 directly after the experiment asks the participants to rate the methods in terms of ease of use (Q4a, variable a) and whether they expect reasonable, realistic and useful results (Q4b, variable c).

Name: _____

Aufgabe 4: Fragen zur gesamten Fallstudie

In welchen Teil der Fallstudienbeschreibung hast du während der Aufgaben 1, 2 oder 3 hineingesehen?

	Aufgabe 1	Aufgabe 2	Aufgabe 3
Ziel des Experiments			
Aufgabenbeschreibung			
Fallstudienbeschreibung Umfeld			
Liste der neun Anforderungen			
Anlage: Funktionale Anforderungen			

Welche für die Abschätzungen nötigen Informationen fehlten dir in der Fallstudienbeschreibung? Welche Annahmen hast du über diese Größen getroffen?

a. Welche der Schätzmethoden waren leicht durchzuführen und welche fielen dir schwer?

	Sehr leicht	leicht	weiß nicht	schwierig	sehr schwierig
– Aufgabe 1:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
– Aufgabe 2:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
– Aufgabe 3:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Begründungen und Kommentare:

b. Wo hast du das Gefühl, sinnvolle, realistische und praktisch verwendbare Bewertungen erzeugt zu haben, und welchen traust du weniger?

	Sehr sinnvoll, Realistisch und brauchbar	einigermaßen sinnvoll, etc.	weiß nicht	eher nicht	gar nicht
– Aufgabe 1:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
– Aufgabe 2:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
– Aufgabe 3:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Begründungen und Kommentare:

c. Wie transparent erschienen dir die Methoden?

	Sehr transparent	Eher transparent	weiß nicht	eher intransparent	sehr intransparent
– Aufgabe 1:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
– Aufgabe 2:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
– Aufgabe 3:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Begründungen und Kommentare:

A.b.6 Questionnaire Q5

Explanation for the reader of this report: Questionnaire Q5: One week after the experiment, during the post-test session, each participant receives a table with his/ her priorities resulting from each method. They are asked which method's results reflects their opinion best (Q5a, variable d). Question Q5b asked whether the damage estimation was easier in method 2 (where absolute values are estimated) or 3 (relative values), or equal. Q5c asked how certain the participants feel concerning their estimations in each method. Question Q5d offered four statistics of frequencies of security incidents and sources of attack taken from the CSI/FBI Computer Crime and Security Survey (Richardson, 2003). On their basis, the reference risk probabilities of two security misuse cases were re-estimated (the participants could look up their former estimations if they wanted to). The participants were asked by Q5e whether and how the statistics facilitated the estimations. Question Q5f showed the resulting priorities of all participants, including averages and standard deviations and asked what might be causes of these deviations.

Questionnaire Q5 was personalized for each participant, containing his/ her own results at Question Q5a.

Aufgabe 5: Diskussion der Ergebnisse

Fragebogen für Teilnehmer Nummer

Bewertung deiner eigenen Ergebnisse

Im Folgenden bewertest du zunächst deine eigenen Ergebnisse.

a. Die folgende Tabelle stellt deine aus den Aufgaben 1 bis 3 resultierenden Prioritäten der Anforderungen dar: Die 1 steht für die wichtigste Anforderung und die 9 für die unwichtigste.

	Aufgabe 1	Aufgabe 2	Aufgabe 3
A1: Benutzeroberfläche übersichtlich und intuitiv gestalten			
A2: Support			
A3: Ähnlichkeit mit einem echten Flohmarkt			
A4: Inspektion der Spezifikationsdokumente			
A5: Verschlüsselte Speicherung der Kundendaten			
A6: Schnelle Hard- und Software			
A7: Standard			
A8: Automatisierte Benachrichtigung der Service-Mitarbeiter bei Ausfall des Systems			
A9: Ersatz-Server			

Das Ergebnis welcher der drei Aufgaben erscheint dir insgesamt am plausibelsten und

warum?

Wie erklärst du dir (falls vorhanden) die Unterschiede zwischen deinen Ergebnissen aus verschiedenen Methoden?

- Liegt es an fehlenden Informationen?
- An Missverständnissen?
- Daran, dass die Schätzungen in Einzelschritte aufgeteilt sind und das Ergebnis der eigenen Abschätzungen nicht vorhersehbar?
- Daran, dass Risikoabschätzungen allgemein schwierig sind?
- Dass das Prinzip des Referenzsystems unklar war oder schwierig anzuwenden?
- Daran, dass sich im Verlauf des Experiments das Wissen über das System und sein Umfeld änderte?
- Andere/ weitere Gründe?

- b. Du hast sowohl in Aufgabe 2 als auch in Aufgabe 3 Schäden abgeschätzt, die durch Misuse Cases verursacht werden. In Aufgabe 2 hast du relativ zu einem Gesamtschaden von 550.000€ geschätzt und in Aufgabe 3 in Bezug auf den Schaden, der an einem Qualitätsziel verursacht wird. Was fiel dir leichter?

- Schaden-Schätzung in Aufgabe 2
- Schaden-Schätzung in Aufgabe 3
- Beides gleich
- Weiß nicht

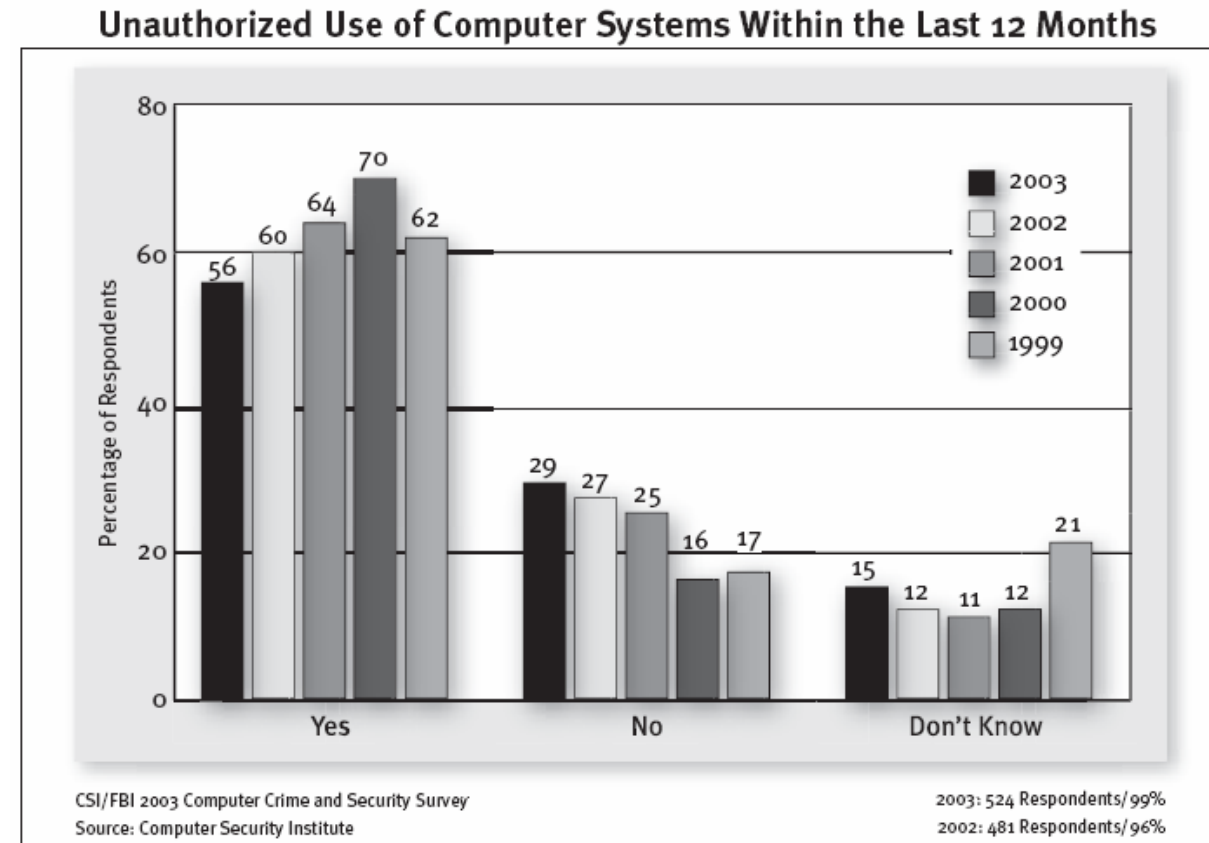
c. Wie sicher warst du dir bei deinen Beurteilungen bei Aufgabe 2 (Restrisiko und Basisrisiko)?

- Sehr sicher
- Eher sicher
- Gemischt
- Eher unsicher
- Sehr unsicher

Wie sicher warst du dir bei deinen Beurteilungen bei Aufgabe 3 (Schätzungen ab Geschäftsschaden bis zum Restrisiko)?

- Sehr sicher
- Eher sicher
- Gemischt
- Eher unsicher
- Sehr unsicher

- d. Im Folgenden siehst du einige Statistiken¹ zu Sicherheitsvorfällen. Gib an, wie hoch du jetzt die Wahrscheinlichkeit für Misuse Case 6 und 7 schätzt.



¹ Quelle: Richardson, Robert: 2003 CSI/FBI Computer Crime and Security Survey. Computer Security Institute. 2003; http://i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2003.pdf (2003)

How Many Incidents?

By percentage (%)	1 to 5	6 to 10	11 to 30	31 to 60	Over 60	Don't Know
2003	38	20	more:16	0	0	26
2002	42	20	8	2	5	23
2001	33	24	5	1	5	31
2000	33	23	5	2	6	31
1999	34	22	7	2	5	29

2003: 356 Respondents/67%, 2002: 321 Respondents/64%, 2001: 348 Respondents/65%, 2000: 392 Respondents/61%, 1999: 327 Respondents/63%

How Many From the Outside?

By percentage (%)	1 to 5	6 to 10	11 to 30	31 to 60	Over 60	Don't Know
2003	46	10	13	0	0	31
2002	49	14	5	0	4	27
2001	41	14	3	1	3	39
2000	39	11	2	2	4	42
1999	43	8	5	1	3	39

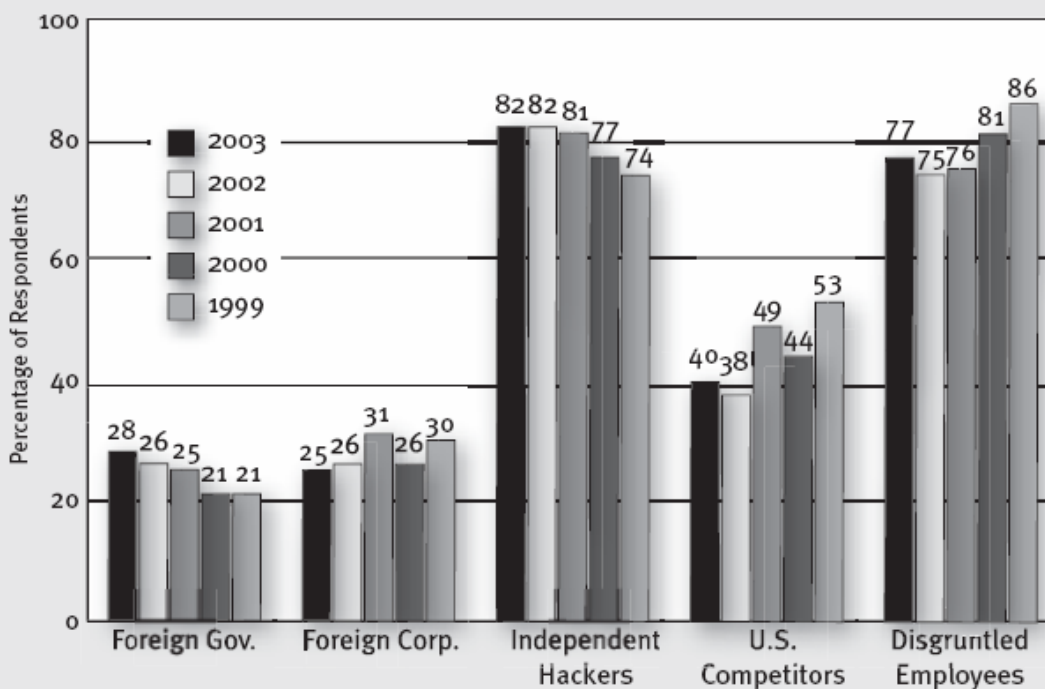
2003: 336 Respondents/63%, 2002: 301 Respondents/60%, 2001: 316 Respondents/59%, 2000: 341 Respondents/53%, 1999: 280 Respondents/54%

How Many From the Inside?

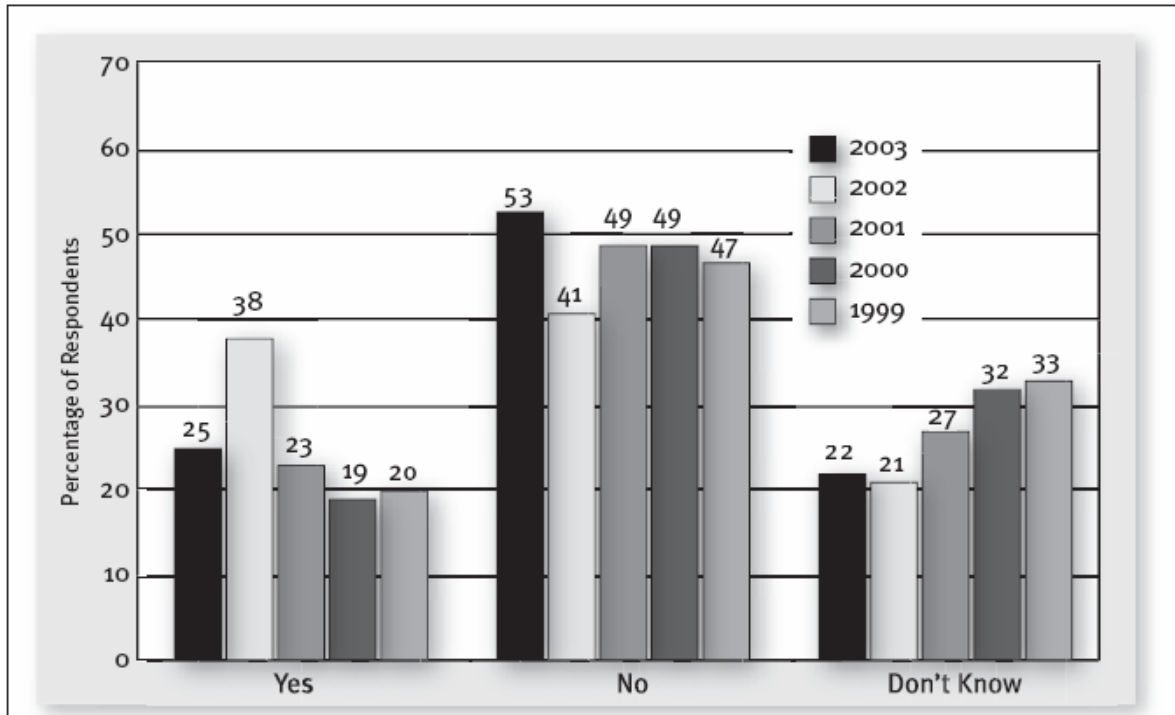
By percentage (%)	1 to 5	6 to 10	11 to 30	31 to 60	Over 60	Don't Know
2003*	45	11	12	0	0	33
2002	42	13	6	2	1	35
2001	40	12	3	0	4	41
2000	38	16	5	1	3	37
1999	37	16	9	1	2	35

2003: 328 Respondents/62%, 2002: 289 Respondents/57%, 2001: 348 Respondents/65%, 2000: 392 Respondents/61%, 1999: 327 Respondents/63%

Likely Sources of Attack



Has Your WWW Site Suffered Unauthorized Access or Misuse Within the Last 12 Months?



Wie hoch schätzt du nun die folgenden beiden Wahrscheinlichkeiten?

Wahrscheinlichkeit für MUC6: „Kundendaten werden von nicht autorisierter Person gelesen“ im Referenzsystem (Restrisiko) in % (auf zwei Jahre gerechnet): _____

Wahrscheinlichkeit für MUC7: „Hacker manipulieren die Seite samt ihrer Inhalte“ im Referenzsystem (Restrisiko) in % (auf zwei Jahre gerechnet): _____

e. Haben die Statistiken die Schätzung erleichtert oder erschwert? Wie?

Vergleich der Ergebnisse der verschiedenen Teilnehmer/innen

f. Die folgenden Tabellen zeigen die Prioritäten, die die verschiedenen Teilnehmer/innen jeweils vergeben haben.

In Aufgabe 1:

	Teiln. 1	2	3	4	5	6	7	8	9	10	Mittelwert	Prio	Quadrat der Standard-abweichung	Standardabweichung
1	1	4	1	4	1	6	8	3	1	2	3,10	1	5,877777778	2,42441287
2	5	9	4	7	2	8	7	5	5	2	5,40	6	5,6	2,36643191
3	2	1	6	2	3	9	2	1	4	7	3,70	2	7,566666667	2,75075747
4	4	2	9	8	8	7	9	2	3	8	6,00	7-8	8,444444444	2,90593263
5	3	3	3	1	4	1	6	7	6	5	3,90	3	4,322222222	2,07899548
6	9	8	8	3	5	5	5	4	9	4	6,00	7-8	5,111111111	2,26077666
7	6	6	5	9	2	3	4	9	7	2	5,30	5	6,677777778	2,58413966
8	7	7	2	5	6	2	3	8	2	6	4,80	4	5,511111111	2,34757558
9	8	5	7	6	9	4	1	6	8	9	6,30	9	6,233333333	2,49666444
														2,46840963

in Aufgabe 2:

	Teiln. 1	2	3	4	5	6	7	8	9	10	Mittelwert	Prio	Quadrat der Standard-abweichung	Standardabweichung
1	3	1	2	3	9	6	4,5	3	3	2	3,65	2	5,447222222	2,333928496
2	9	5	5	5	3	1	8	5	6	5	5,20	5	5,066666667	2,250925735
3	7	6	9	9	6,5	7	2	8	8	9	7,15	9	4,447222222	2,108843812
4	8	7	8	4	2	3	4,5	9	4	7	5,65	6	5,891666667	2,427275565
5	5	2	1	1	1	2	7	2	9	1	3,10	1	8,322222222	2,884826203
6	1	8	6	8	6,5	8	3	7	5	5	5,75	7	5,402777778	2,324387613
7	4	3	7	2	4	9	1	6	7	3	4,60	4	6,488888889	2,547329757
8	6	9	3,5	6,5	6,5	4,5	9	4	1	8	5,80	8	6,622222222	2,573367875
9	2	4	3,5	6,5	6,5	4,5	6	1	2	6	4,20	3	4,177777778	2,043961296
														2,388316261

in Aufgabe 3:

	Teiln. 1	2	3	4	5	6	7	8	9	10	Mittelwert	Prio	Quadrat der Standard-abweichung	Standardabweichung
1	6	2	4	5	9	8	5	3	5	3	5,00	5	4,88888889	2,21108319
2	8	5	6	8	2	1	8	6	9	7	6,00	6	7,111111111	2,66666667
3	7	6	9	6	6,5	5	4	9	8	9	6,95	8	3,136111111	1,77090686
4	9	7	7	7	3	7	9	8	6	8	7,10	9	2,98888889	1,72884033
5	5	3	1	1	1	2	3	2	7	1	2,60	1	4,044444444	2,01108042
6	4	9	5	9	6,5	6	6	7	3	5	6,05	7	3,802777778	1,95007122
7	3	4	8	4	4	9	1	5	4	2	4,40	4	6,044444444	2,45854519
8	2	8	2,5	2,5	6,5	3,5	7	4	1	6	4,30	3	5,78888889	2,40601099
9	1	1	2,5	2,5	6,5	3,5	2	1	2	4	2,60	1	2,933333333	1,71269768
														2,10176695

Wie erklärst du dir diese Unterschiede?

c) Spreadsheet tables

The tables on questionnaires Q1, Q2 and Q3 were also implemented as spreadsheet tables, which sometimes contained additional columns. The participants' results were entered in these tables and additional calculations made, like calculation of misuse case risk and of countermeasure benefit.

The table for Q1 (method 1) was exactly as in the document above. The tables supporting method 2 and 3 are given in the following sections.

Tables for Method 2, Questionnaire Q2

Misuse Case	Wahrscheinlichkeit p des Misuse Case in %	Schaden anteilig am Nutzen des Geschäftsziels (550.000€), in %	Schaden d in €	$p \times d =$ Restrisiko in €
MUC1: Benutzerfehler vereitelt geplanten Kauf			0	0
MUC2: Vernachlässigung von Übersichtlichkeitsanforderungen bei der Softwareentwicklung führt zu Kundenverlust			0	0
MUC3: Verkäufer gibt nach längerer Zeit entnervt auf, einen Artikel einstellen zu wollen			0	0
MUC4: Benutzer ohne technischem Hintergrund verstehen technische Begriffe/ Oberfläche nicht -> lange Lernphase und Kundenverlust			0	0
MUC5: Programm hilft Benutzer nicht bei Fehleingaben			0	0
MUC6: Kundendaten werden von nicht autorisierter Person gelesen			0	0
MUC7: Hacker manipulieren die Seite samt ihrer Inhalte			0	0
MUC8: Ineffizienz durch lange Antwortzeiten des Systems			0	0
MUC9: Komplexität des Systems führt zu hohem Wartungsaufwand			0	0
MUC10: geringe Benutzereffizienz durch schlechtes Auffinden der Information, Nichterkennen von Wesentlichem			0	0

MUC11: Code kann bei Änderungen im Systemumfeld nicht oder nur aufwändig wiederverwendet werden			0	0
MUC12: Der einzige Server fällt aus, erst nach Stunden bemerkt ein Service-Mitarbeiter den Ausfall durch Zufall; ein Ersatz-Server ist nicht vorhanden; die Benutzer können tagelang nicht auf die Seite			0	0

Misuse Case	Nicht realisierte Gegenmaßnahme	Wahrscheinlichkeit p des Misuse Case in %	Schaden relativ zum Restschaden in Tab. 2.1, in %	Schaden d in €	$p \times d =$ Basisrisiko in €
MUC1: Benutzerfehler vereitelt geplanten Kauf	A2			0	0
MUC1: Benutzerfehler vereitelt geplanten Kauf	A1			0	0
MUC2: Vernachlässigung von Übersichtlichkeitsanforderungen bei der Softwareentwicklung führt zu Kundenverlust	A1			0	0
MUC3: Verkäufer gibt nach längerer Zeit entnervt auf, einen Artikel einstellen zu wollen	A2			0	0
MUC4: Benutzer ohne technischem Hintergrund verstehen technische Begriffe/ Oberfläche nicht - > lange Lernphase und Kundenverlust	A3			0	0
MUC5: Programm hilft Benutzer nicht bei Fehleingaben	A4			0	0
MUC6: Kundendaten werden von nicht autorisierter Person gelesen	A5			0	0
MUC7: Hacker manipulieren die Seite samt ihrer Inhalte	A5			0	0
MUC8: Ineffizienz durch lange Antwortzeiten des Systems	A6			0	0
MUC9: Komplexität des Systems führt zu hohem Wartungsaufwand	A7			0	0
MUC10: geringe Benutzereffizienz durch schlechtes Auffinden der Information, Nichterkennen von Wesentlichem	A7			0	0
MUC11: Code kann bei Änderungen im Systemumfeld nicht oder nur aufwändig wiederverwendet werden	A7			0	0
MUC12: Der einzige Server fällt aus, erst nach Stunden bemerkt ein Service-Mitarbeiter den Ausfall durch Zufall; ein Ersatz-Server ist nicht vorhanden; die Benutzer können tagelang nicht auf die Seite	A8			0	0
MUC12: Der einzige Server fällt aus, erst nach Stunden bemerkt ein Service-Mitarbeiter den Ausfall durch Zufall; ein Ersatz-Server ist nicht vorhanden; die Benutzer können tagelang nicht auf die Seite	A9			0	0

Gegenmaßnahme = Anforderung	Bezug zu welchem Misuse Case	Nutzen (in €) in Bezug auf Misuse Case = Basisrisiko - Restrisiko	Gesamtnutzen der Gegenmaßnahme

A1: Benutzeroberfläche übersichtlich und intuitiv gestalten, z.B. durch aussagekräftige Beschriftungen, Befolgen von Usability-Richtlinien	MUC1	0	0
“	MUC2	0	”
A2: Support	MUC1	0	0
“	MUC3	0	”
A3: Ähnlichkeit mit einem echten Flohmarkt	MUC4	0	0
A4: Inspektion der Spezifikationsdokumente	MUC5	0	0
A5: Verschlüsselte Speicherung der Kundendaten	MUC6	0	0
“	MUC7	0	”
A6: Schnelle Hard- und Software	MUC8	0	0
A7: Standard	MUC9	0	0
“	MUC10	0	”
“	MUC11	0	”
A8: Automatisierte Benachrichtigung der Service-Mitarbeiter bei Ausfall des Systems	MUC12	0	0
A9: Ersatz-Server	MUC12	0	0

Tables for Method 3, Questionnaire Q2

Geschäftsschaden	Anteil des Geschäftsziels "Profit", der durch den jeweiligen Geschäftsschaden zerstört wird, in %	Unsicherheit des Anteils (+/- Prozentpunkte)	Qualitätsmangel	Wahrscheinlichkeit, dass der Qualitätsmangel zum Geschäftsschaden führt, in %	Schaden des Qualitätsmangels in €
Geringer Marktanteil, schlechter Ruf			Usability-Mängel		0
Missbrauch des Systems			Sicherheitsmängel		0
Ausfall des Systems			Sicherheitsmängel		0
			hoher Wartungsaufwand		0
			Unzuverlässigkeit der Hardware/ Software		0
Hohe Betriebskosten			hoher Wartungsaufwand		0
			Performanzmängel		0

Qualitätsziel	Qualitätsmängel	relativer Schaden des nichterreichten Qualitätsziels am Qualitätsmangel, in %	Nutzen des Qualitätsziels in €
Übersichtlichkeit der Benutzeroberfläche	Usability-Mängel		0
Erlernbarkeit der Benutzer-Tasks	Usability-Mängel		0
Fehlertoleranz der Benutzeroberfläche	Usability-Mängel		0
Effizienz	Usability-Mängel		0
Effizienz	Performanzmängel		0
Wartbarkeit	hoher Wartungsaufwand		0
Portierbarkeit	hoher Wartungsaufwand		0
Safety, d.h. Ausfallsicherheit	Unzuverlässigkeit der HW/SW		0
Wiederherstellbarkeit	Unzuverlässigkeit der HW/SW		0
Safety, d.h. Ausfallsicherheit	Sicherheitsmängel		0
Security, d.h. Missbrauchsicherheit	Sicherheitsmängel		0
Vertraulichkeit der Kundendaten	Sicherheitsmängel		0
Integrität der Daten und Prozesse	Sicherheitsmängel		0

Qualitätsziel	Misuse Case	Wahrscheinlichkeit p des Misuse Case in %	Schadenanteilig am Qualitätsziel in %	Schaden d in €	$p \times d =$ Restrisiko in €
Übersichtlichkeit der Benutzeroberfläche	MUC2: Vernachlässigung von Übersichtlichkeitsanforderungen bei der Softwareentwicklung führt zu Kundenverlust	0		0	0
Erlernbarkeit der Benutzer-Tasks	MUC4: Benutzer ohne technischem Hintergrund verstehen Begriffe/ Oberfläche nicht -> lange Lernphase und Kundenverlust	0		0	0
Fehlertoleranz der Benutzeroberfläche	MUC1: Benutzerfehler vereitelt geplanten Kauf	0		0	0
Fehlertoleranz der Benutzeroberfläche	MUC5: Programm hilft Benutzer nicht bei Fehleingaben	0		0	0
Effizienz	MUC3: Verkäufer gibt nach längerer Zeit entnervt auf, einen Artikel einstellen zu wollen	0		0	0
Effizienz	MUC8: Ineffizienz durch lange Antwortzeiten des Systems	0		0	0
Effizienz	MUC10: geringe Benutzereffizienz durch schlechtes Auffinden der Information, Nichterkennen von Wesentlichem	0		0	0
Portierbarkeit	MUC11: Code kann bei Änderungen im Systemumfeld nicht oder nur aufwändig wiederverwendet werden	0		0	0
Safety = Ausfallsicherheit	MUC12: Der einzige Server fällt aus, erst nach Stunden bemerkt ein Service-Mitarbeiter den Ausfall durch Zufall; ein Ersatz-Server ist nicht vorhanden; die Benutzer können tagelang nicht auf die Seite	0		0	0
Wartbarkeit	MUC9: Komplexität des Systems führt zu hohem Wartungsaufwand	0		0	0
Vertraulichkeit der Kundendaten	MUC6: Kundendaten werden von nicht autorisierter Person gelesen	0		0	0
Integrität der Daten und Prozesse	MUC5: Programm hilft Benutzer nicht bei Fehleingaben	0		0	0
Integrität der Daten und Prozesse	MUC7: Hacker manipulieren die Seite samt ihrer Inhalte	0		0	0

Misuse Case	Nicht realisierte Gegenmaßnahme	Wahrscheinlichkeit p des Misuse Case in %	Schaden im Verhältnis zum Restschaden in Tab. 3.3	Schaden d in €	$p \times d =$ Basisrisiko in €
MUC1: Benutzerfehler vereitelt geplanten Kauf	A2	0	0	0	0
MUC1: Benutzerfehler vereitelt geplanten Kauf	A1	0	0	0	0
MUC2: Vernachlässigung von Übersichtlichkeitsanforderungen bei der Softwareentwicklung führt zu Kundenverlust	A1	0	0	0	0
MUC3: Verkäufer gibt nach längerer Zeit entnervt auf, einen Artikel einstellen zu wollen	A2	0	0	0	0
MUC4: Benutzer ohne technischem Hintergrund verstehen Begriffe/ Oberfläche nicht -> lange Lernphase und Kundenverlust	A3	0	0	0	0
MUC5: Programm hilft Benutzer nicht bei Fehleingaben	A4	0	0	0	0
MUC6: Kundendaten werden von nicht autorisierter Person gelesen	A5	0	0	0	0
MUC7: Hacker manipulieren die Seite samt ihrer Inhalte	A5	0	0	0	0
MUC8: Ineffizienz durch lange Antwortzeiten des Systems	A6	0	0	0	0
MUC9: Komplexität des Systems führt zu hohem Wartungsaufwand	A7	0	0	0	0
MUC10: geringe Benutzereffizienz durch schlechtes Auffinden der Information, Nichterkennen von Wesentlichem	A7	0	0	0	0
MUC11: Code kann bei Änderungen im Systemumfeld nicht oder nur aufwändig wiederverwendet werden	A7	0	0	0	0
MUC12: Der einzige Server fällt aus, erst nach Stunden bemerkt ein Service-Mitarbeiter den Ausfall durch Zufall; ein Ersatz-Server ist nicht vorhanden; die Benutzer können tagelang nicht auf die Seite	A8	0	0	0	0
MUC12: Der einzige Server fällt aus, erst nach Stunden bemerkt ein Service-Mitarbeiter den Ausfall durch Zufall; ein Ersatz-Server ist nicht vorhanden; die Benutzer können tagelang nicht auf die Seite	A9	0	0	0	0

Gegenmaßnahme = Anforderung	Bezug zu welchem Misuse Case	Nutzen in Bezug auf Misuse Case = Basisrisiko - Restrisiko	Gesamt-nutzen der Gegenmaßnahme
A1: Benutzeroberfläche übersichtlich und intuitiv gestalten, z.B. durch aussagekräftige Beschriftungen, Befolgen von Usability-Richtlinien	MUC1	0	0
“	MUC2	0	“
A2: Support	MUC1	0	0
“	MUC3	0	“
A3: Ähnlichkeit mit einem echten Flohmarkt	MUC4	0	0
A4: Inspektion der Spezifikationsdokumente	MUC5	0	0
A5: Verschlüsselte Speicherung der Kundendaten	MUC6	0	0
“	MUC7	0	“
A6: Schnelle Hard- und Software	MUC8	0	0
A7: Standard	MUC9	0	0
“	MUC10	0	“
“	MUC11	0	“
A8: Automatisierte Benachrichtigung der Service-Mitarbeiter bei Ausfall des Systems	MUC12	0	0
A9: Ersatz-Server	MUC12	0	0

A.2 Experiment 2

a) Presentation slides

Experiment: Priorisierung von Anforderungen

4. Juli 2007



Institut für Informatik
Neuenheimer Feld 326
D-69120 Heidelberg, Germany
<http://www-swe.informatik.uni-heidelberg.de>
herrmann@informatik.uni-heidelberg.de



RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG



Thema des Experiments

Experiment

- ▶ Einführung
- Zeitplan
- Methode 1
- Methode 2

- **Priorisierung von Anforderungen = Grundlage von Entscheidungen zwischen Anforderungen, z.B.**
 - **Projektmanagement-Entscheidungen**
 - **Technische Entscheidungen**

The topic of the experiment is the prioritization of requirements, what is the basis for decisions among requirements, e.g. project management decisions or technical decisions.

Einführung: mögl. Priorisierungskriterien

Experiment

- ▶ Einführung
- Zeitplan
- Methode 1
- Methode 2

- Nutzen
 - Kosten & (Kalender) Zeit
 - Bedeutung der Quelle
 - Risiko, Fehleranfälligkeit
 - Dringlichkeit, Sanktion
 - Auswirkungen, Komplexität, Machbarkeit
 - Volatilität
- } Nutzen/ Kosten
} Nutzen - Kosten

As the participants of experiment 2 (unlike those of experiment 1) had not heard about requirements prioritization before during the lecture, here we give a very short introduction to requirements prioritization in general. This slide lists some possible prioritization criteria.

Einführung: Prioritätenskalen

Experiment

- ▶ Einführung
- Zeitplan
- Methode 1
- Methode 2

- **Muss/ wichtig/ unwichtig oder 3/ 2/ 1**
- **Sortierung: 1., 2., 3., ...**
- **Bewertung von Nutzen oder Kosten in einer Geldwährung oder anderem Maß**

Scales of priorities can be must/ important/ not important or 3/2/1, requirements can be ranked or the benefit or cost can be estimated in a currency or any other measure.

Ziel des Experiments

Experiment

- ▶ Einführung
- Zeitplan
- Methode 1
- Methode 2

**Test von Methoden für die risikobasierte
Priorisierung von Anforderungen:**

- Nutzen quantitative Risikoabschätzungen?

The objective of the experiment is to test methods for the risk based prioritization of requirements. Our question is whether quantitative risk estimations are useful.

Methoden im Experiment

Experiment

- ▶ Einführung
- Zeitplan
- Methode 1
- Methode 2

Methoden:

1. Intuitive Sortierung
2. Risikoabschätzung in MOQARE

The two methods compared in the experiment are intuitive ranking and risk estimation in MOQARE.

Methode 1: intuitive Sortierung

- Experiment
- ▶ Einführung
 - Zeitplan
 - Methode 1
 - Methode 2

1. Intuitive Sortierung in zwei Schritten

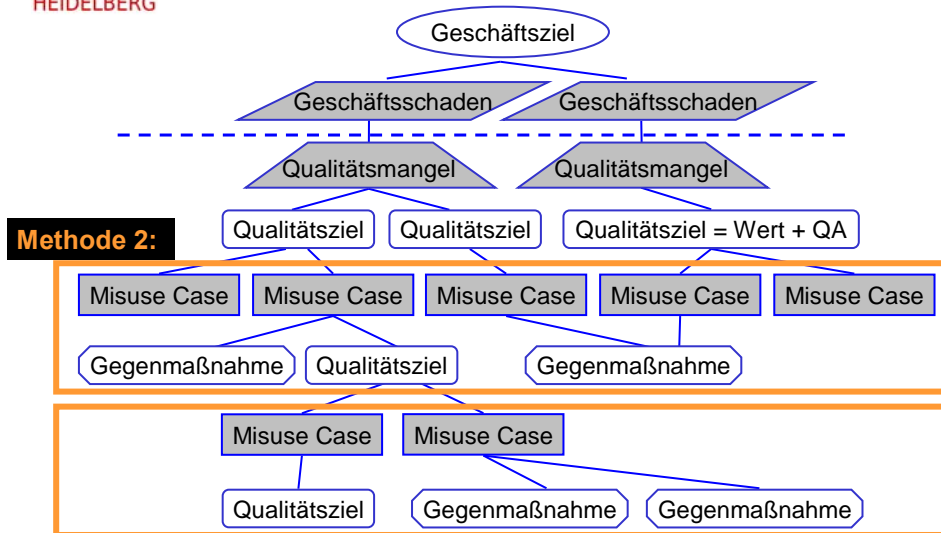
- Grobsortierung
- Feinsortierung

Anforderung (Gegenmaßnahme)	Grobbewertung: "hoher Nutzen", "durchschnittlich nützlich", "geringer Nutzen"	Feinbewertung
A1	<i>hoch</i>	1. (1 = wichtigste)
A2	<i>gering</i>	6
A3	<i>gering</i>	5

Folie 7
rungen

The intuitive ranking is done in two steps: coarse-grained evaluation in terms of “high benefit”, “average benefit” or “low benefit”, and then the requirements are ranked in their order, number 1 being the most important requirement.

Priorisierung in MOQARE



Methode 2:

Folie 8

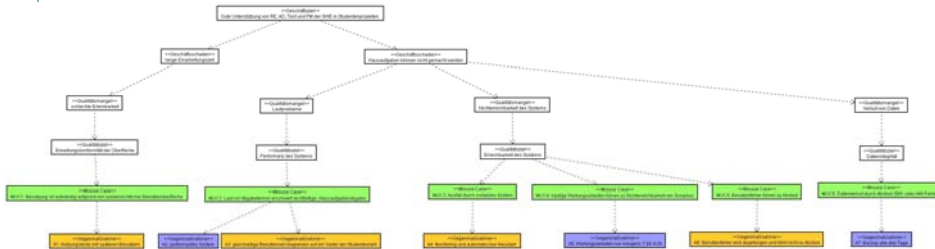
Method 2 derives the countermeasure benefit from misuse case risk, i.e. it refers to the lower part of the MOQARE Misuse Tree (which the experiment participants already know from the lecture).

Zu betrachtendes System: Sysiphus

Sysiphus in der Lehre

Experiment

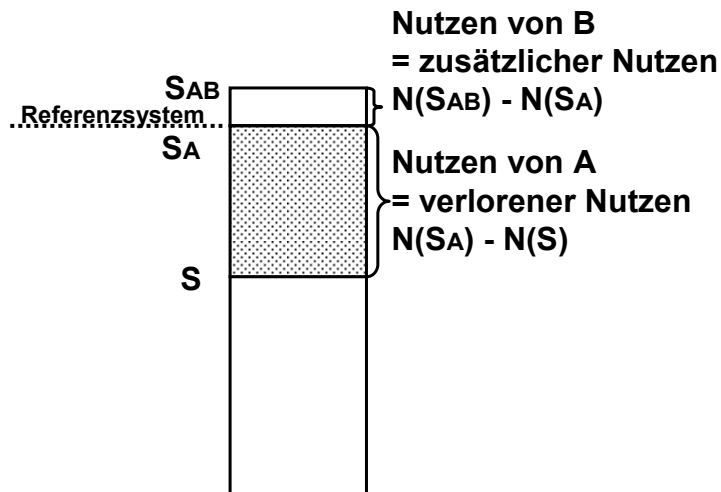
Einführung



Siehe: Experiment2_MisuseTree.jpg

This slide refers to the homework which the students had to make in preparation for the experiment. They had identified possible misuse cases and countermeasures for Sysiphus. This slide explains those misuse cases and countermeasures which will be treated during the experiment.

Referenzsystem = Status Quo



The reference system is the status quo, i.e. Sysiphus as it is currently implemented and was used by the experiment participants. For a countermeasure B, which is not implemented in the reference system, the benefit measures the benefit which is added by implementing B, additionally to the reference system. For a countermeasure A, which is implemented in the reference system, the benefit measures the benefit which is lost by not implementing A.

Zeitplan: Teams 1-4

Experiment

▶ Einführung	14:10-14:30	Einführung
▶ Zeitplan		
▶ Methode 1	14:35-14:55	Methode 1
▶ Methode 2	14:55-15:00	Fragebogen 1
	15:00-15:05	Priorisierung in MOQARE
	15:05-15:45	Methode 2
	15:45-15:50	Fragebögen 2 und 3

This is the time plan for teams 1-4, which started with method 1 and then executed method 2.

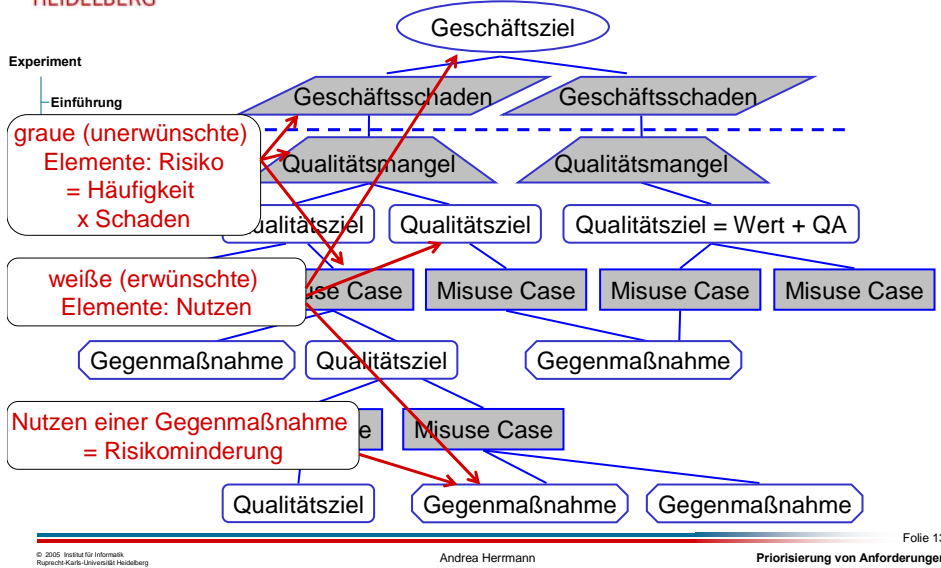
Zeitplan: Teams 5-8

Experiment

▶ Einführung		
▶ Zeitplan	15:55-16:15	Einführung
▶ Methode 1	16:15-16:20	Priorisierung in MOQARE
▶ Methode 2	16:25-17:05	Methode 2
	17:05-17:10	Fragebogen 2
	17:10-17:30	Methode 1
	17:30-17:35	Fragebögen 1 und 3

This is the time plan for teams 5-8 (remark: There was no team 7), who started with method 2 and then executed method 1.

Priorisierung in MOQARE



Explanation of the principle of how countermeasures are prioritized in MOQARE: grey (unwanted) elements are prioritized by their risk, which is defined as the product of probability and caused damage. White (wanted) elements are prioritized according to their benefit. A countermeasure adds benefit by reducing misuse case risk, and therefore its benefit is equal to the risk reduction.

Methode 2: Definitionen

Wir betrachten einen Zeitraum von **einem Monat**; während der Vorlesungszeit und bei normaler gleichzeitiger Benutzung durch SWE I und SWE IIa Studenten.

Experiment
Einführung
Zeitplan
Methode 1
Methode 2

Häufigkeit p : beschreibt, **wie oft** ein Misuse Case **pro Monat** und pro Person eintritt ($p=1$ für „1 Mal pro Monat“)

Schaden d : beschreibt den Schaden, der **durchschnittlich** bei **jedem Eintritt** des Misuse Case entsteht, gemessen in **verlorener Kalenderzeit in Stunden**, die einem Benutzer verloren geht.

Folglich beschreibt das **Risiko $p \times d$** den **Schaden**, der durch diesen Misuse Case **während eines Monats im Mittel** entsteht.

*Beispiel: Ein Student arbeitet um Mitternacht an Sysphus, aber es stürzt ab. Es wird am nächsten Morgen um 7 Uhr erfolgreich neu gestartet, aber der Student arbeitet erst um 9 Uhr weiter. Dann sind 9 Stunden verloren gegangen, in denen nicht gearbeitet werden konnte, obwohl die verlorene Arbeitszeit insgesamt vielleicht nur 20 Minuten beträgt.
Wenn dies zwei Mal pro Monat passiert, beträgt das Risiko $2 \times 9h = 18h$.*

© 2005 Institut für Informatik, Ruprecht-Karls-Universität Heidelberg
Andrea Herrmann
Folie 14
Priorisierung von Anforderungen

Some definitions had to be made for method 2, which – as not questionnaires were handed out to the participants – in experiment 2 were defined on the slides. These definitions are: The period of time is one month, during lecture time and normal and concurrent use by students of the SWE I and SWE IIa courses. (These are the courses which the experiment participants attended.)

Probability p describes how often a misuse case happens per month and per person. $p=1$ means “once per month”.

Damage d describes which is caused in average by each occurrence of the misuse case, measured in lost calendar time in hours, which are lost to a user.

Consequently, the risk $p \times d$ describes the damage which is caused by this misuse case per month, in average.

Example: A student works with Sysiphus at midnight, but it crashes. It is re-booted successfully the next morning at 7 o'clock, but the student does not continue to work before 9 o'clock. Then, 9 hours have been lost, during which no work was possible, although the lost working time maybe is only 20 minutes. If this happens twice per month, the risk is $2 \times 9h = 18h$.



Vielen Dank für eure Teilnahme!

Experiment

— Einführung
— Zeitplan
— Methode 1
— Methode 2

Andrea Herrmann

**Institut für Informatik
Neuenheimer Feld 326
D-69120 Heidelberg
Germany**

**<http://www-swe.informatik.uni-heidelberg.de>
herrmann@informatik.uni-heidelberg.de**

Thank you for your participation!

b) Material: Questionnaires

Additionally to performing the estimations in the group (based on the spreadsheet tables shown in the following section), the participants individually answered to questions of three questionnaires:

- Questionnaire Q1 evaluates method 1
- Questionnaire Q2 evaluates method 2
- Questionnaire Q3 compares method 1 with method 2

In the following sections, the original text of the questionnaires is given, after an explanation for the reader of this report, set in italic.

Questionnaire Q1

Explanation for the reader of this report: Questionnaire Q1 evaluates method 1 and was answered directly after the execution of method 1. It asked:

- *Question 1.1: How certain are you that the priorities resulting from the group discussion are realistic? (variable c)*
- *1.2: Were there requirements for which you are especially uncertain that they are classified right? If yes: Which? (the list of requirements was offered here) Why? Which information was missing? (variable f)*
- *1.3: Do you think that you have been involved adequately during the discussions and that your proposals and objections have been considered sufficiently?*
- *1.4 Comments*

Experiment zur Anforderungspriorisierung:

Name: _____

Bewertung der Methode 1 **(intuitive Sortierung: Grob- und Feinsortierung)**

1.1 Wie sicher bist du dir, dass die in der Gruppe abgestimmten Prioritäten realistisch sind?

- Sehr sicher
- Eher sicher
- Gemischt
- Eher unsicher
- Sehr unsicher

*1.2 Gab es Anforderungen, bei denen du besonders **unsicher** bist, dass sie richtig eingeordnet wurden? Wenn ja: Welche?*

- A1: Nutzungstests mit späteren Benutzern
- A2: Performantes System (mehr Hauptspeicher & schnellere Prozessoren & effizientere Algorithmen)
- A3: gleichzeitige Benutzerzahl begrenzen auf ein Viertel der Studentenzahl
- A4: Monitoring und automatischer Neustart
- A5: Wartungsarbeiten nur morgens 7:00-9:00
- A6: Benutzerfehler wird abgefangen und führt nicht zu Absturz
- A7: Backup alle drei Tage

Warum? Welche Informationen haben dir eventuell gefehlt?

1.3 Bist du der Meinung, dass du bei der Diskussion während Methode 1 angemessen beteiligt warst und deine Vorschläge und Einwände genügend berücksichtigt wurden?

- Vollständig
- Eher ja
- Teilweise
- Eher nein
- Gar nicht

Falls nein: Bei welchen Anforderungen nicht?

- A1: Nutzungstests mit späteren Benutzern
- A2: Performantes System (mehr Hauptspeicher & schnellere Prozessoren & effizientere Algorithmen)
- A3: gleichzeitige Benutzerzahl begrenzen auf ein Viertel der Studentenzahl
- A4: Monitoring und automatischer Neustart
- A5: Wartungsarbeiten nur morgens 7:00-9:00
- A6: Benutzerfehler wird abgefangen und führt nicht zu Absturz
- A7: Backup alle drei Tage

1.4 Kommentare hierzu:

Questionnaire Q2

Explanation for the reader of this report: Questionnaire Q2 evaluates method 2 and was answered directly after the execution of method 2.

- 2.1: *How certain are you that the probability and damage estimations resulting from the group discussion are realistic? (variable c)*
- 2.2: *How certain are you that the priorities resulting from the group discussion are realistic? (variable c)*
- 2.3: *What do you think how uncertain are the group's estimations of the probabilities? (If an estimation $p = 2$ times per month presumably lies between 1.5 and 2.5 times, then the accuracy is " ± 0.5 times".) (variable e)*
- 2.4: *Were there misuse cases for which you are especially uncertain? If yes: Which? (the list of misuse cases was offered here) (variable f)*
- 2.5: *What do you think how uncertain are the group's estimations of the damages? (If an estimation of 4 hours presumably lies between 2 and 4 hours, then the accuracy is " ± 2 hours".) (variable e)*
- 2.6: *Were there misuse cases for which you are especially uncertain? If yes: Which? (the list of misuse cases was offered here) (variable f)*
- 2.7: *Do you think that you have been involved adequately during the discussions and that your proposals and objections have been considered sufficiently?*
- 2.8 *Comments*

Experiment zur Anforderungspriorisierung:

Name: _____

Bewertung der Methode 2 **(Schätzung von Wahrscheinlichkeit und Schaden der Misuse Cases)**

2.1 Wie sicher bist du dir, dass die in der Gruppe abgestimmten Schätzungen der Wahrscheinlichkeiten und Schäden realistisch sind?

- Sehr sicher
- Eher sicher
- Gemischt
- Eher unsicher
- Sehr unsicher

2.2 Wie sicher bist du dir, dass die resultierenden Prioritäten realistisch sind?

- Sehr sicher
- Eher sicher
- Gemischt
- Eher unsicher
- Sehr unsicher

2.3 Für wie genau hältst du die in der Gruppe ermittelten Schätzungen der Häufigkeiten? (Liegt ein Schätzwert $p = 2$ Mal pro Monat vermutlich zwischen 1,5 und 2,5 Mal, beträgt die Genauigkeit $\pm 0,5$ Mal.)

2.4 Wenn du bei einzelnen Misuse Cases besonders unsicher bist, kreuze diese an.

- MUC1: Benutzung ist aufwändig aufgrund von unübersichtlicher Benutzeroberfläche
- MUC2: Last vor Abgabetermin erschwert rechtzeitige Hausaufgabenabgabe
- MUC3: Ausfall durch instabiles System
- MUC4: häufige Wartungsarbeiten führen zu Nichterreichbarkeit von Sysiphus
- MUC5: Benutzerfehler führen zu Absturz
- MUC6: Datenverlust durch Absturz (SW- oder HW-Fehler)

2.5 Für wie genau hältst du die in der Gruppe ermittelten Schätzungen der Schäden? (Liegt ein Schätzwert von 4 Stunden vermutlich zwischen 2 und 6, sind das ± 2 Stunden.)

2.6 Wenn du bei einzelnen Misuse Cases besonders unsicher bist, kreuze diese an.

- MUC1: Benutzung ist aufwändig aufgrund von unübersichtlicher Benutzeroberfläche
- MUC2: Last vor Abgabetermin erschwert rechtzeitige Hausaufgabenabgabe
- MUC3: Ausfall durch instabiles System
- MUC4: häufige Wartungsarbeiten führen zu Nichterreichbarkeit von Sysiphus
- MUC5: Benutzerfehler führen zu Absturz
- MUC6: Datenverlust durch Absturz (SW- oder HW-Fehler)

2.7 Bist du der Meinung, dass du bei der Diskussion während Methode 2 angemessen beteiligt warst und deine Vorschläge und Einwände genügend berücksichtigt wurden?

- Vollständig
- Eher ja
- Teilweise
- Eher nein
- Gar nicht

Falls nein: Bei wie vielen der insgesamt 28 geschätzten Werte nicht?

2.8 Kommentare hierzu:

Questionnaire Q3

Explanation for the reader of this report: Questionnaire Q3 compared method 1 with method 2 and was answered after the execution of both methods and after Q1 and Q2.

- 3.1: Which of the methods were easy to execute and which was difficult? (variable b)
- 3.2: Explanations and comments
- 3.3: Which of the methods would you presumably have found easy or difficult, if you had executed it alone?
- 3.4: Explanations and comments
- 3.5: How have the group discussion been useful for the results?
- 3.6: Did this discussion also have disadvantages?
- 3.7: Which were the advantages and disadvantages of the quantitative risk estimation compared to the intuitive ranking of the requirements?
- 3.8: Compare the priorities resulting from both methods. Do they reflect your view? (variable d)
- 3.9: Explanations and comments

Experiment zur Anforderungspriorisierung:

Name: _____

Vergleich der Methoden

3.1 Welche der Methoden waren leicht durchzuführen und welche war schwierig?

	Sehr leicht	leicht	weiß nicht	schwierig	sehr schwierig
Methode 1:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Methode 2:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

3.2 Begründungen und Kommentare:

3.3 Welche der Methoden hättest du vermutlich leicht oder schwierig gefunden, wenn du sie alleine durchgeführt hättest?

	Sehr leicht	leicht	weiß nicht	schwierig	sehr schwierig
Methode 1:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Methode 2:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

3.4 Begründungen und Kommentare:

3.5 Welchen Nutzen brachte die Diskussion in der Gruppe für die Ergebnisse?

3.6 Brachte diese Diskussion auch Nachteile?

3.7 Welche Vor- und Nachteile hatte die quantitative Abschätzung der Risiken gegenüber der intuitiven Sortierung der Anforderungen?

3.8 Vergleiche die Prioritäten, die aus beiden Methoden resultieren. Spiegeln sie deine Meinung wider?

	Sehr gut	eher ja	teilweise	eher nein	sehr schlecht
Methode 1:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Methode 2:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

3.9 Begründungen und Kommentare:

c) Spreadsheet tables

In experiment 2, the spreadsheet tables were used during the experiment, to document the results of the group discussions.

Table for method 1

Explanations which are given on top of the table:

1. Schritt: Grobbewertung ausfüllen für jede Anforderung, in Bezug auf das Geschäftsziel "gute Unterstützung von RE, AD, Test und PM der SWE in Studentenprojekten"; der Misuse Case dient der Illustration.

Gelb markierte Gegenmaßnahmen sind nicht erfüllt. Hier stellt sich die Frage: Welchen zusätzlichen Nutzen würde die Anforderung bringen, wenn sie erfüllt wäre?

Blau markierte Gegenmaßnahmen sind schon erfüllt. Hier stellt sich die Frage: Welche Nutzen würde man verlieren (bzw. Schaden erfahren), wenn die Anforderung nicht erfüllt wäre?

2. Schritt: Feinbewertung: innerhalb jeder Gruppe die einzelnen Anforderungen so bewerten, dass ein Gesamt-Ranking entsteht

Wähle die Zahl 1 für die wichtigste und 8 für die unwichtigste Anforderung.

Für jeden Eintrag sollte nicht länger als 2 Minuten gebraucht werden. Lieber unklare Punkte überspringen und später darauf zurück kommen, wenn noch Zeit bleibt. Insgesamt stehen 15 Minuten zur Verfügung.

Anforderung/Gegenmaßnahme	Grobbewertung: "hoher Nutzen", "durchschnittlich nützlich", "geringer Nutzen"	Feinbewertung	Misuse Case
A1: Nutzungstests mit späteren Benutzern			MUC1: Benutzung ist aufwändig aufgrund von unübersichtlicher Benutzeroberfläche
A2: Performantes System (mehr Hauptspeicher & schnellere Prozessoren & effizientere Algorithmen)			MUC2: Last vor Abgabetermin erschwert rechtzeitige Hausaufgabenabgabe
A3: gleichzeitige Benutzerzahl begrenzen auf ein Viertel der Studentenzahl			MUC2: Last vor Abgabetermin erschwert rechtzeitige Hausaufgabenabgabe
A4: Monitoring und automatischer Neustart			MUC3: Ausfall durch instabiles System
A5: Wartungsarbeiten nur morgens 7:00-9:00			MUC4: häufige Wartungsarbeiten führen zu Nichterreichbarkeit von Sysiphus
A6: Benutzerfehler wird abgefangen und führt nicht zu Absturz			MUC5: Benutzerfehler führen zu Absturz
A7: Backup alle drei Tage			MUC6: Datenverlust durch Absturz (SW- oder HW-Fehler)

Tables for method 2

Estimation of the reference risk:

Trotz aller Gegenmaßnahmen besteht für die meisten Misuse Cases auch im Referenzsystem noch ein Restrisiko.

Schätzt das Restrisiko pro Misuse Case im Referenzsystem. Das Referenzsystem ist das aktuell verwendete System.

Schätzt hierzu für jeden Misuse Case die durchschnittliche Häufigkeit p für das Auftreten in einem Monat pro Person, sowie den erwarteten Schaden d in verlorenen Kalenderstunden (nicht Arbeitsstunden).

Das Risiko ist definiert als das Produkt aus Häufigkeit und Schaden $p \times d$ und wird von der Tabelle berechnet.

Für jeden Eintrag sollte nicht länger als 2 Minuten gebraucht werden. Lieber unklare Punkte überspringen und später darauf zurück kommen, wenn noch Zeit bleibt. Geplante Zeit: 20 Minuten.

Misuse Case	umgesetzte Gegenmaßnahme	durchschnittliche Häufigkeit p des Misuse Case	Schaden d in verlorenen Kalenderstunden	$p \times d =$ Restrisiko in verlorenen Kalenderstunden pro Monat
MUC1: Benutzung ist aufwändig aufgrund von unübersichtlicher Benutzeroberfläche				0
MUC2: Last vor Abgabetermin erschwert rechtzeitige Hausaufgabenabgabe	A2: Performantes System (mehr Hauptspeicher & schnellere Prozessoren & effizientere Algorithmen)			0
MUC3: Ausfall durch instabiles System				0
MUC4: häufige Wartungsarbeiten führen zu Nichterreichbarkeit von Sysiphus	A5: Wartungsarbeiten nur morgens 7:00-9:00			0
MUC5: Benutzerfehler führen zu Absturz				0
MUC6: Datenverlust durch Absturz (SW- oder HW-Fehler)	A7: Backup alle drei Tage			0

Estimation of the varied risk:

Um später den Nutzen der Gegenmaßnahmen zu ermitteln, wird nun das Basisrisiko abgeschätzt. Das Basisrisiko misst das Risiko, wenn eine bestimmte Gegenmaßnahme nicht oder zusätzlich zum Referenzsystem realisiert wird.

Um jeweils das Basisrisiko abzuschätzen, geht in der folgenden Tabelle zeilenweise vor. Stellt euch vor, im Gegensatz zum Referenzsystem sei die eine angegebene Gegenmaßnahme nicht (blau) oder zusätzlich (gelb) implementiert.

Wie hoch sind dann jeweils durchschnittliche Häufigkeit und Schaden des angegebenen Misuse Case? Die Werte des Restrisikos sind als Vergleichswerte eingetragen und sollen nicht überschrieben werden.

Das Risiko $p \times d$ wird von der Tabelle berechnet.

Für jeden Eintrag sollte nicht länger als 2 Minuten gebraucht werden. Lieber unklare Punkte überspringen und später darauf zurück kommen, wenn noch Zeit bleibt. Geplante Zeit: 20 Minuten.

Misuse Case	Nicht oder zusätzlich realisierte Gegenmaßnahme	durchschnittliche Häufigkeit p (Restrisiko zu Herrmann: beide Spalten nicht)	Schaden d (Restrisiko zum Vergleich)	durchschnittliche Häufigkeit p des Misuse Case (Basisrisiko)	Schaden d in verlorenen Kalenderstunden (Basisrisiko)	$p \times d =$ Basisrisiko in verlorenen Kalenderstunden
MUC1: Benutzung ist aufwändig aufgrund von unübersichtlicher Benutzeroberfläche	A1: Nutzungstests mit späteren Benutzern	0	0			0
MUC2: Last vor Abgabetermin erschwert rechtzeitige Hausaufgabenabgabe	A2: Performantes System (mehr Hauptspeicher & schnellere Prozessoren & effizientere Algorithmen)	0	0			0
MUC2: Last vor Abgabetermin erschwert rechtzeitige Hausaufgabenabgabe	A3: gleichzeitige Benutzerzahl begrenzen auf ein Viertel der Studentenzahl	0	0			0
MUC3: Ausfall durch instabiles System	A4: Monitoring und automatischer Neustart	0	0			0
MUC4: häufige Wartungsarbeiten führen zu Nichterreichbarkeit von Sysiphus	A5: Wartungsarbeiten nur morgens 7:00-9:00	0	0			0
MUC5: Benutzerfehler führen zu Absturz	A6: Benutzerfehler wird abgefangen und führt nicht zu Absturz	0	0			0
MUC6: Datenverlust durch Absturz (SW- oder HW-Fehler)	A7: Backup alle drei Tage	0	0			0

Calculation of countermeasure benefit and ranking of the countermeasures

Der Nutzen der Gegenmaßnahme wird automatisch berechnet, die Priorität muss von Hand ermittelt werden. (The countermeasure benefit is calculated automatically, the priority must be defined manually.)

Gegenmaßnahme = Anforderung	Bezug zu welchem Misuse Case	Nutzen der Gegenmaßnahme (in Kalender-tagen) in Bezug auf Misuse Case = Basisrisiko - Restrisiko	Priorität
A1: Nutzungstests mit späteren Benutzern	MUC1: Benutzung ist aufwändig aufgrund von unübersichtlicher Benutzeroberfläche	0	
A2: Performantes System (mehr Hauptspeicher & schnellere Prozessoren & effizientere Algorithmen)	MUC2: Last vor Abgabetermin führt zu Absturz	0	
A3: gleichzeitige Benutzerzahl begrenzen auf ein Viertel der Studentenzahl	MUC2: Last vor Abgabetermin führt zu Absturz	0	
A4: Monitoring und automatischer Neustart	MUC3: Ausfall durch instabiles System	0	
A5: Wartungsarbeiten nur morgens 7:00-9:00	MUC4: häufige Wartungsarbeiten führen zu Nichterreichbarkeit von Sysiphus	0	
A6: Benutzerfehler wird abgefangen und führt nicht zu Absturz	MUC5: Benutzerfehler führen zu Absturz	0	
A7: Backup alle drei Tage	MUC6: Datenverlust durch Absturz (SW- oder HW-Fehler)	0	

Annex B: Data and Data Analysis

This annex for experiment 1 and 2 describes the quantitative analysis of the variables defined in section 3. For each variable, the results from both experiments and all methods are presented together. Their interpretation, especially how we believe that these variables have been influenced by the influencing factors, is discussed in section 6.

A.1 Time Consumption

Method 1 demands to determine only two values (group and priority) for each of the countermeasures. Method 2 demands the estimation of two probabilities and two damages per countermeasure. Method 3 additionally to the probability and damage estimations as in method 2, one value for each business damage, quality deficiency and quality goal must be estimated.

In experiment 1, the time needed for risk estimation is significantly higher than for the ranking. The average time needed for Q1 was 6.6 minutes for those 7 participants who noted it. Q2 took an average of 31.8 min and Q3 was 18.3 min. The time consumption of method 3 was calculated to be 42.8 min because Q3 reuses risk estimations from Q2. In experiment 2, the time need averaged over those groups which performed this method first, was 17.5 minutes for method 1 and 37.3 minutes for method 2. (We count only these groups because of the learning effect observed.)

From these numbers, we calculated the time need per countermeasure (Table 2) and also the time need per estimation (Table 3), as the number of estimations per countermeasure in methods 2 and 3 depends on the number of misuse cases.

Table 2: Time need in minutes per countermeasure

	Experiment 1	Experiment 2
Method 1	0.73	2.50
Method 2	3.53	5.33
Method 3	4.76	--

Table 3: Time need in minutes per estimation

	Experiment 1	Experiment 2
Method 1	0.37	1.25
Method 2	0.61	1.43
Method 3	0.56	--

A.2 Priorities

The resulting **priorities** of the countermeasures varied widely among the participants and groups in both experiments, for all three methods, as can be seen from Table 13 to

Table 17. This means that they differ greatly about the importance of the countermeasures. The same countermeasure could have the highest priority (1) for one participant/ group and the lowest for another. This was the case even in method 1, where the results were transparent to and manipulable by the estimators, while in method 2 and 3 the lacking transparency and indirect manipulability of the priorities could possibly lead to results which are unexpected by the estimators.

In method 1, the averages of the priorities lie between 3.0 and 6.4 for the individual countermeasures. In method 2, the averages are between 3.1 and 7.2; in method 3, between 2.6 and 7.1. In experiment 2, all groups agreed that R3 is one of the least important: In method 1, all seven groups gave R3 the lowest priority 7, while in method 2, this was the case for five groups, once it received priority 6 and once 5. The priority averages in method 1 (not counting R3), vary from 2.57 to 5.00, and in method 2 from 2.29 to 4.21.

We are sure that these wide ranges are not caused by the misunderstanding whether “1” stands for the highest priority or the lowest. In method 1, the priority 1 countermeasure for all participants was found in the “high benefit” group. In methods 2 and 3, the priorities were determined by us, based on the calculated benefits.

A.a Standard Deviation of Priorities

The **standard deviations** s found among the ten estimations of the nine priorities in experiment 1 are shown in Table 4 and among the seven estimations of the seven priorities in experiment 2 in Table 5. The differences between the standard deviations between the methods during the same experiment are very low and statistically not significant. The standard deviations in experiment 2 are lower than in experiment 1 because fewer countermeasures were prioritized, but also when divided by the average priority (which is $(n+1)/2$), in experiment 2 the standard deviation is lower (see Table 6).

Table 4. Variable a: The standard deviations s found among the priorities of the ten participants in experiment 1 for each single countermeasure: \bar{s} denotes the average of s over all participants, s_{\min} the minimum value found for any countermeasure, and s_{\max} the maximum.

	\bar{s}	s_{\min}	s_{\max}
Method 1	2.35	2.01	2.84
Method 2	2.38	2.04	2.88
Method 3	2.10	1.71	2.67

Table 5. Variable a: standard deviation of the priorities of the seven groups in experiment 2 (calculated for each countermeasure, then averaged over all countermeasures) (* was 0 for countermeasure R7)

\bar{s}	s_{\min}	s_{\max}	
1.40528	1.25357*	1.951800	Method 1
1.58414	0.78680 (R7) resp. 1.0965	2.64575	Method 2

Table 6. Variable a: coefficient of variation (=standard deviation/ average = \bar{s} / \bar{p}) of the priorities p

	Experiment 1	Experiment 2
Method 1	0.470	0.351
Method 2	0.476	0.396
Method 3	0.420	--

A.b Ease of Use

The **ease of use** of each method as rated by the participants is shown in Table 7. We attribute points to the answers: Very easy =2 points, Easy =1, Undecided =0, Difficult =-1, Very difficult =-2.

Table 7. Variable b: The ease of use as assessed by the participants (in experiment 1: Q4a; in experiment 2: Question 3.1). Given are the average values in points, averaged over all participants

	Experiment 1	Experiment 2
Method 1	1.0	1.12 Points
Method 2	-1.2	-0.92 Points
Method 3	-0.2	---

A.c Participants Expect their Estimations to Be Realistic

In experiment 1, immediately after the estimations, but before they knew the resulting priorities, we asked the participants whether they expected to have made reasonable, realistic and useful estimations (question Q4b, see Table 8). In experiment 2, they were asked whether they believe that their results were realistic. At this point of time, they knew the priorities (questions 1.1, 2.1 and 2.2). Points were attributed to the answers: very = 2 points, rather =1, undecided =0, rather not=-1, not at all = -2 points.

It is interesting to note that in experiment 2, the average points for the probability and damage estimations were 0.13, i.e. lower than for the priorities (but not statistically significantly due to high variations). This means that some participants trusted in the method to deliver priorities which are more realistic than the risk estimations on which they are based.

Table 8. Variable c: Do the participants expect the priorities to be realistic? Given are the average values, averaged over all participants, in points

	Experiment 1	Experiment 2
Method 1	1.00	1.04
Method 2	0.10	0.17 points
Method 3	0.50	---

A.d Accuracy of the Results

Accuracy here means that after the experiment the participants considered the countermeasures priorities which resulted from their risk estimations to be plausible and reflecting their views. In experiment 1, this question was asked in the post-test session one week after the experiment, in experiment 2, the question was asked during the experiment session directly after the estimations.

In experiment 1, nine participants answered question Q5a (qualitative answers). Four were in favour of method 1, two wrote, that their first impression was that method 1 delivered the most plausible results, because they corresponded to their intuitive priorities, but as method 3 was a systematic method, probably this method should provide the best results. In one case, the priorities obtained by methods 2 and 3 were almost equal, therefore there was no difference. Another participant wrote that all methods delivered plausible as well as less plausible results for the different countermeasures. The last participant wrote that they were all plausible: Method 1 reflected his own perception, neglecting risk and cost. Method 2 and 3 were plausible as well, taking into account risk and cost

In experiment 2 (Question 3.8 und 3.9), method 1 got an average of 1.29 points as a result to this question; method 2 got 0.13 points on a scale between -2 and +2 (very well = 2 points, rather yes=1, undecided =0, rather not=-1, very badly= -2 points). This difference is statistically significant.

A.e Uncertainty of the Estimated Values and of Countermeasure Benefits

The participants individually were asked how uncertain they think their results are. In experiment 1, the uncertainty of the probability or damage estimations were circa $\pm 10\%$. In experiment 2 (questions 2.3 and 2.5), the uncertainties were estimated about 40%.

A.f Frequency of Naming a Misuse Case or Countermeasure as Especially Uncertain

The participants were asked how uncertain they expect their estimations to be (variable f). They also were asked to name the countermeasures or misuse cases for which they considered their estimated values especially uncertain. We counted how often a certain countermeasure (in method 1) or a certain misuse case (in method 2) was named. To compare the methods and countermeasures, we calculated the average frequency with which a countermeasure or misuse case was named here, averaged over all countermeasures/ misuse cases. We did so, because the number of countermeasures and misuse cases was not the same in the two experiments and the methods applied. The results are summarized in Table 9.

In experiment 1, regarding method 1, R8 was never named and R3 only once. R5 was mentioned twice, R1, R2, R4 and R9 three times, R7 four times and R6 five times by the ten participants. Concerning method 2, eight out of ten participants named specific misuse cases, while six additionally or instead said that they were uncertain about practically all of them. If we take the latter group literally, each misuse case was named with the average frequency 0.7. If we do not count the participants stating they were uncertain concerning all misuse cases, the numbers are about 0.2 for the probability estimations as well as for the damage estimations.

In experiment 2, the questions 1.2, 2.4 and 2.6 helped to investigate this question. In question 1.2 regarding method 1, R1 was named seven times (by the 18 participants who

answered to this question), R6 and R7 four times, R2, R4 and R5 three times and R3 only once. Except for R1, which seems to be especially difficult to judge, and R3, which caused almost no irritation, most countermeasures seem to be equally difficult. Few correlations are seen among the answers of members of the same group. Only once, all three group members agreed that R1 was difficult, and three times two of three or four group members agreed about the same countermeasure.

In question 2.4 concerning the probability estimation in method 2, misuse cases #1-3 were named four times (by the 18 participants who answered this question), misuse case #4 only once, misuse case #5 nine times and misuse case #6 eight times. Only once, all three group members agreed about the same misuse case: for misuse case #5 they all found probability estimation difficult. Five times, two group members agreed about the same misuse case.

In question 2.6 concerning the damage estimation in method 2, the misuse cases were named with the following frequencies (by the 18 participants who answered this question): misuse case #6 nine times, #1 and #4 seven times each, #3 five times, #5 four times, and #2 only once. Only once all three group members agreed concerning #3. Seven times, two group members considered the same misuse case's damage estimation to be difficult. This means an average frequency for a misuse case of 0.3. These numbers are approximately the same as for the probability estimation.

The differences observed between methods (during the same experiment) and during different experiments for the same method are statistically significant.

Table 9: variable f: average frequency with which a certain countermeasure or misuse case was named as being difficult to estimate per participant (* marks the results we obtain when the six participants who said that they were uncertain for all misuse cases are taken literally)

	Experiment 1*	Experiment 1	Experiment 2
Method 1		0.27	0.20
Method 2, probability estimation	0.73	0.24	0.28
Method 2, damage estimation	0.67	0.20	0.31

A.g Participants Feel Certain

How the participants rated their **certainty** about their estimations in experiment 1 is shown in Table 10. We attribute points to the answers and in the right column give the average rate of each method, averaged over all answers. The participants felt more certain with method 1 than with both the others. We do not decide whether they feel more certain about method 2 than about method 3, as the difference between method 2 and 3 was caused by the vote of one person out of 10 and is not statistically relevant.

Table 10. variable g: How certain did the participants feel concerning their estimations? No-one chose the options "very uncertain" or "very certain" (Q5c)

	Rather uncertain	Partly certain	Rather certain	Average

	(-1 point)	(0 point)	(1 point)	
Method 1		3	7	0.7
Method 2	3	6	1	-0.2
Method 3	2	6	2	0.0

A.3 Influence of Statistics

In experiment 1, we tested the influence of public statistics provided to the estimators. In Q5d, 8 out of 9 participants now clearly attributed different probabilities (reference risk) to the two security misuse cases, usually much higher ones, see Table 11 and Table 12. One participant wrote that he estimated the same probabilities as before (0.5% and 0.1%), but we doubt whether they were really derived from the statistics, as they differ too much from the estimations of the other participants.

To question Q5e (whether the statistics facilitated the probability estimation), we received qualitative answers. Some of them were: “They definitely were helpful. One feels more certain, thanks to this information.” Others were more sceptical: “I would say that the statistics have strongly influenced my estimations. However, I wonder how similar the systems of these companies are to the reference system in the case study. Only when this is known, one can say whether the high estimated value is justified. I think that I still do not have enough information to deliver a good estimation.” All together, 4 participants out of 9 wrote that the statistics were helpful, while one wrote they did not influence the estimation and four wrote they influenced the estimations, but they still were sceptical whether the estimated values were exact.

Table 11. Probability estimations for misuse case #6 („Customer data are read by an unauthorized person“)

	without statistics (Q2)	with statistics (Q5d)
Average over all participants	7.64%	45.3%
Standard deviation	17.2%	20.3%
Coefficient of variation = standard dev./ average	2.25	0.45

Table 12. Probability estimations for misuse case #7 (“Hackers manipulate the flea market including ist content“)

	without statistics (Q2)	with statistics (Q5d)
Average over all participants	15.0%	33.1%
Standard deviation	21.8%	21.7%
Coefficient of variation = standard dev./ average	1.45	0.66

We expected that when providing the participants with several statistics, the standard deviation (relative to the average) of their estimations to decrease (variable a). As one can see in Table 11 and Table 12, the estimated probabilities still differed among the participants. This was to be expected because four statistics were given, which did not apply to exactly the

same environment as the case study. Therefore, interpretations and adaptations were necessary. The coefficient of variation became less than half by using the statistics.

Table 13: Priorities resulting from experiment 1 with method 1 (“1” standing for the most important one): The “1” is row “R1” and column “1” means that according to the results of participant 1, countermeasure R1 is the most important one. Column “Average” is the average priority of a countermeasure, averaged over all participants, and column “Priority” shows the order of priority for these averages.

Participant:	1	2	3	4	5	6	7	8	9	10	Average	Priority	Standard deviation
R1	1	4	1	4	1	6	8	3	1	1	3.00	1	2.49443826
R2	5	9	4	7	2	8	7	5	5	4	5.60	5	2.11869981
R3	2	1	6	2	3	9	2	1	4	3	3.30	2	2.49666444
R4	4	2	9	8	8	7	9	2	3	7	5.90	6	2.84604989
R5	3	3	3	1	4	1	6	7	6	2	3.60	3	2.11869981
R6	9	8	8	3	5	5	5	4	9	5	6.10	7	2.18326972
R7	6	6	5	9	7	3	4	9	7	8	6.40	9	2.01108042
R8	7	7	2	5	6	2	3	8	2	6	4.80	4	2.34757558
R9	8	5	7	6	9	4	1	6	8	9	6.30	8	2.49666444

Table 14: Priorities resulting from experiment 1 with method 2 (“1” indicating the most important one)

Participant:	1	2	3	4	5	6	7	8	9	10	Average	Priority	Standard deviation
R1	3	1	2	3	9	6	4,5	3	3	2	3.65	2	2.333928496
R2	9	5	5	5	3	1	8	5	6	5	5.20	5	2.250925735
R3	7	6	9	9	6,5	7	2	8	8	9	7.15	9	2.108843812
R4	8	7	8	4	2	3	4,5	9	4	7	5.65	6	2.427275565
R5	5	2	1	1	1	2	7	2	9	1	3.10	1	2.884826203
R6	1	8	6	8	6,5	8	3	7	5	5	5.75	7	2.324387613
R7	4	3	7	2	4	9	1	6	7	3	4.60	4	2.547329757
R8	6	9	3,5	6,5	6,5	4,5	9	4	1	8	5.80	8	2.573367875
R9	2	4	3,5	6,5	6,5	4,5	6	1	2	6	4.20	3	2.043961296

Table 15: Priorities resulting from experiment 1 with method 3 (“1” indicating the most important one)

Participant:	1	2	3	4	5	6	7	8	9	10	Average	Priority	Standard deviation
R1	6	2	4	5	9	8	5	3	5	3	5.00	5	2.21108319
R2	8	5	6	8	2	1	8	6	9	7	6.00	6	2.66666667
R3	7	6	9	6	6,5	5	4	9	8	9	6.95	8	1.77090686
R4	9	7	7	7	3	7	9	8	6	8	7.10	9	1.72884033
R5	5	3	1	1	1	2	3	2	7	1	2.60	1	2.01108042
R6	4	9	5	9	6,5	6	6	7	3	5	6.05	7	1.95007122
R7	3	4	8	4	4	9	1	5	4	2	4.40	4	2.45854519
R8	2	8	2,5	2,5	6,5	3,5	7	4	1	6	4.30	3	2.40601099
R9	1	1	2,5	2,5	6,5	3,5	2	1	2	4	2.60	1	1.71269768

Table 16: Priorities resulting from experiment 2 with method 1 (“1” indicating the most important one)

Team	1	2	3	4	5	7	8	Average	Priority	Standard deviation
R1	5	2	5	5	4	1	1	3.29	3-4	1.88982237
R2	4	4	3	2	3	4	6	3.71	5	1.25356634
R3	7	7	7	7	7	7	7	7.00	7	0
R4	1	5	2	1	2	2	5	2.57	1	1.71824939
R5	2	6	6	6	5	6	4	5.00	6	1.52752523
R6	3	1	4	3	6	3	3	3.29	3-4	1.49602648
R7	6	3	1	4	1	5	2	3.14	2	1.95180015

Table 17: Priorities resulting from experiment 2 with method 2 (“1” indicating the most important one)

Team	1	2	3	4	5	7	8	Average	Priority	Standard deviation
R1	2.5	5	4	4	5	3	6	4.21	6	1.219875091
R2	6	2	3	6	4	4	2	3.86	3	1.676163420
R3	7	7	6	7	7	7	5	6.57	7	0.786795792
R4	4	1	2	1	2	2	4	2.29	1	1.253566341
R5	2.5	4	5	5	3	5	3	3.93	4	1.096531328
R6	1	3	7	3	6	1	7	4.00	5	2.645751311
R7	5	6	1	2	1	6	1	3.14	2	2.410295378

Copyright 2005, Software
All rights reserved.

This publication, whether the whole or part of the material is concerned, may not be commercially photocopied, reproduced, distributed in electronic, mechanical or any other form, stored in data bases or translated without previous written permission from the authors. For the reproduction or distribution of the publication for private purposes no written consent is required.