Experimental Validation of a Risk Assessment Method

Eelco Vriezekolk[1,3(✉)], Sandro Etalle[2,3], and Roel Wieringa[3]

[1] Radiocommunications Agency Netherlands, Groningen, Netherlands
[2] Eindhoven University of Technology, Eindhoven, Netherlands
[3] University of Twente, Enschede, Netherlands
eelco.vriezekolk@agentschaptelecom.nl

Abstract. [**Context and motivation**] It is desirable that requirement engineering methods are reliable, that is, that methods can be repeated with the same results. Risk assessments methods, however, often have low reliability when they identify risk mitigations for a system based on expert judgement. [**Question/problem**] Our goal is to assess the reliability of an availability risk assessment method for telecom infrastructures, and to identify possibilities for improvement of its reliability. [**Principal ideas/results**] We propose an experimental validation of reliability, and report on its application. We give a detailed analysis of sources of variation, explain how we controlled them and validated their mitigations, and motivate the statistical procedure used to analyse the outcome. [**Contribution**] Our results can be used to improve the reliability of risk assessment methods. Our approach to validating reliability can be useful for the assessment of the reliability of other methods.

Keywords: Reliability · Risk assessment · Expert judgement · Experiment design · Telecommunications

1 Introduction

Risk assessment is the identification, analysis, and evaluation of risks, and provides the arguments for the choice and justification of risk mitigations [12]. It can be viewed as a way to transform high-level system goals (to avoid risks) into more detailed system requirements (to implement specific mitigations). For example, in security risk assessment, the high-level system goals of confidentiality and availability can be transformed into the more detailed system requirements that there be a firewall and that a function must be implemented redundantly.

Risks assessments are often performed by experts who assess (that is: identify, analyse and evaluate) risks on the basis of best available expert-knowledge of an architectural model. It is known that such expert judgements may have a low reliability [9]. We call a method reliable if it can be repeated with the same results [27]. Other terms in use for this concept are repeatability, stability, consistency, and reproducibility.

Testing the reliability of a risk assessment method is an important issue, which has however received very little attention in the literature. If a risk

S.A. Fricker and K. Schneider (Eds.): REFSQ 2015, LNCS 9013, pp. 1–16, 2015.
DOI: 10.1007/978-3-319-16101-3_1

assessment method is not quite reliable, then its results will always largely depend on the intuition and the expertise of the expert carrying it out. This weakens the ability of decision makers to justify risk mitigation actions that are based on such assessments.

In previous papers we have illustrated the RASTER method for assessing availability risks in telecom infrastructures [24]. A test of RASTER with real experts in a real assessment task has shown that RASTER can achieve useful results within limited time, but did not provide evidence about its reliability [22,23].

In this paper, we illustrate the method we have developed for validating RASTER's reliability. Our approach is based on an experiment, guided by a general checklist to ensure that all important aspects are adequately addressed [25]. Here, we illustrate the choices we have made and the methodologies we have applied to ensure a scientific assessment. We believe that our approach is sufficiently general to be applicable to other requirements engineering methods as well.

We describe risk assessment, and the RASTER method in particular, briefly in Sect. 2. Our approach to testing reliability of a method is presented in Sect. 3. In Sect. 4 and 5 we describe the design and outcome of an experiment using this approach. We discuss implications of this for practice and for research in Sect. 6.

2 Background and Related Work

In what follows, a telecom operator is a company that provides an infrastructure for communication, such as Deutsche Telekom in Germany. Examples of telecom services provided by an infrastructure are voice calling between two parties, internet access, virtual private networks, etc. End users are companies or individuals that use these services, such as banks, shops, police and emergency services, and citizens.

Nowadays, a typical telecom service uses the physical infrastructure and services of several independent telecom operators. These operators may not be aware that their infrastructure is used for the particular service. For example, an operator of fiber-optic transmission network typically leases capacity to other operators and will therefore not know what end-user services are being offered. The end-user organisations' availability requirements are therefore not (fully) known by the operators. Operators strive for high availability and resilience, but are not able to adapt their network to accommodate the availability requirements of individual end users. For some classes of users these availability requirements are very strong, for example for fire and emergency services. Reliable risk assessments can therefore be very important for telecom services.

To side-step the problem of low reliability of expert judgements, risk assessments are sometimes based on checklists and best practices, in what we call 'compliance-based methods'. These compliance-based methods are not sufficient for today's telecom networks, mainly because of three reasons. First, as explained above, telecom operators aim for local optimisations that may have detrimental effects on global availability. Second, the infrastructure is extremely complex, and composed of fixed and mobile networks, using PSTN, internet, wireless and cable

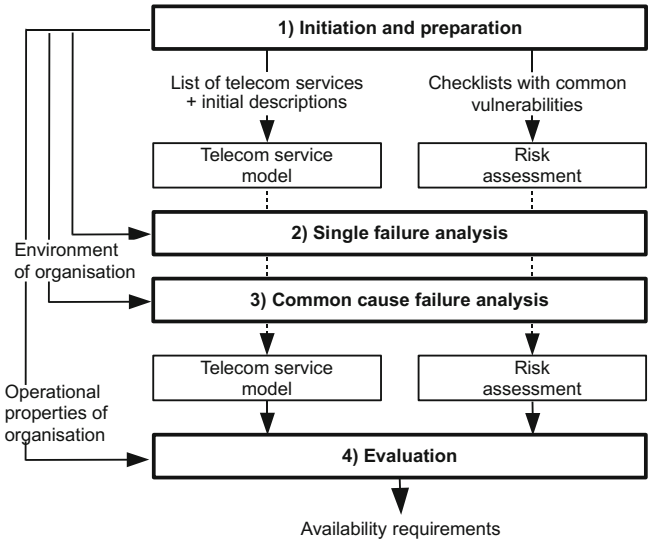

Fig. 1. The four stages of the RASTER method. Stages are shown in bold; documents and flow of information in standard lines.

infrastructures. Third, the infrastructure is in a state of continuous evolution, and threats to the infrastructure evolve as well. This makes compliance-based risk assessments even less effective than risk-based assessments.

Risk assessment methods can be quantitative (e.g. [2,16]) or qualitative (e.g. [3,11]). Quantitative methods estimate probability and impact of risks by ratio scales or interval scales, qualitative methods estimate probability or impact by an ordinal scale, for example 'Low–Medium–High'. Due to lack of information, availability risks for telecom infrastructures have to be estimated qualitatively. This means that expert judgement plays a crucial role. This reduces reliability of risk assessments, either because a single expert makes different estimates for the same object at different times, or because multiple experts make different estimates at the same time. Herrmann *et al.* argue that one reason why reliability is low is that risk estimation requires a lot of information [7,9]. They report that group discussions, although time consuming, have a moderating effect.

The goal of the RASTER method is to guide experts in doing an availability risk assessment on behalf of an end user, that is as reliable as possible, given the constraints of limited information about the target of assessment. The RASTER method is typically executed by a team of analysts comprising of telecom experts as well as domain experts from the end-users organisation [24]. The method consists of four stages (Fig. 1):

1. collect initial information and determine the scope of the analysis, mostly based on existing documentation from the end user and its telecom suppliers. The results include an initial architecture model and a list of common vulnerabilities;

2. analyse single failures (single incidents affecting a single architectural component, e.g. a cable break). After a few iterations, this results in an updated architecture model (see Fig. 2 for an example);
3. analyse common cause failures (single incidents affecting multiple components simultaneously, e.g. a faulty software update that has been applied to multiple routers);
4. perform the final risk evaluations, also considering other stakeholders' interests and possible reactions that may influence the range of acceptable risk mitigations.

Assessments in RASTER are mostly qualitative, using an ordinal scale, and explicitly take into account uncertainty, lack of consensus, and lack of information. Each vulnerability is assessed through two factors: Frequency (indicating likelihood of an incident because of this vulnerability) and Impact (indicating effects, repairability, and number of actors involved). The decision which mitigations to apply is out of scope of RASTER. These decisions are not made by the analysts but by a stakeholder responsible for the mitigation budget, who typically must trade off risk mitigation requirements against other kinds of requirement.

3 Our Approach to Testing Reliability of a Method

We define reliability as repeatability, and so we are interested in how much variation there is across different instances of using the method, where we will control all possible contextual sources of variation in results. We want to understand how to minimise the variation caused by the method itself. Internal causes are inherent to the method, and will be present in any application of the method. For example, the method may be ambiguously described or underspecified. Contextual causes of variation are due to the subject applying the method, the environment during application, or to other aspects of the context in which the method is applied. For example, the time available for application of the method may have been too short. Contextual causes of variation will be present regardless of any particular method being used. We consider a method to be reliable if the variation of the results of using it is small, when contextual causes of variation are held constant.

The sources of variation in an experiment are different from those in the field. We need to control variation in order to be able to draw conclusions from the experiment. Controls therefore only need to be effective within the setting of the experiment; it is not necessary that successful mitigations transfer to field settings.

We now describe our approach of keeping contextual causes of variation constant.

3.1 Controlling Variation

Mitigation of contextual causes of variation involves 1) identification and mitigation of contextual causes, and 2) validation of the effectiveness of mitigations.

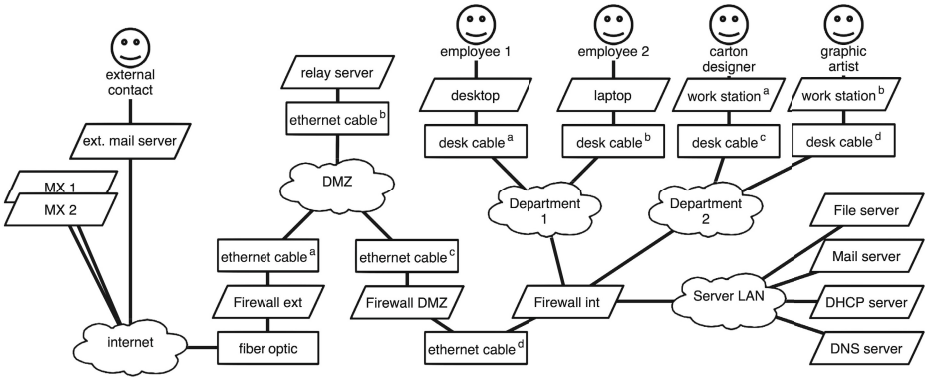

Fig. 2. Example of a telecommunication service model, showing the email service used by a company. Shapes indicate the type of components: equipment (rectangle), cables (parallelogram), and unknown links (cloud). Lines indicate that components are physically connected.

1) Identification and Mitigation. Contextual sources of variation can arise from three areas: a) from the subjects applying the method, b) from the case to which the method is applied, and c) from the circumstances and environment in which the method is applied. In practice it will be impossible to remove contextual causes altogether, but steps can be taken to reduce them, or to measure them so that we can reason about their possible influence on the outcome. Because contextual conditions are controlled, the testing of the reliability of a method will be a laboratory experiment.

a) Subjects Applying the Method. We identified three causes for variation arising from the participants in reliability experiments. Participants may not *understand* the task, not *be able* to perform the task, or not *be willing* to do so.

First, *misapplication and misunderstanding of the method by the participants* can cause variation. If the participants do not have a clear understanding of the method and the task at hand, then they may improvise in unpredictable ways. This can be mitigated by providing a clear and concise case which would be easy to explain, together with clear instructions and reference materials. Furthermore, the clarity of these instructions and the task itself can be tested in a try-out, in which the experiment is conducted with a few participants. Experiences from the try-out can then be used to improve the experiment setup.

Second, *lack of experience and expert knowledge* can cause variation. Even when participants understand the method, they still require skills and knowledge to apply the method properly. Researchers in empirical software engineering often use students as subjects in experiments. It is sometimes claimed that the use of students instead of professionals severely limits the validity of software engineering experiments, because students display behaviour that diverges from that of experts. However, it is valid to use students as a model for experts on the

condition that students are representative with respect to the properties being investigated. Just like a paper model of a house can be used to study some (but not all) properties of a real house, students can be used to study some (but not all) properties of professionals applying software engineering methods [26]. Some studies have indeed found that certain kinds of behaviour are similar across experts and students [10,17,20]. Be this as it may, industry reality often precludes the use of experts in experiments, regardless of how desirable that would be from an experimenter's point of view. Testing with students is cheaper and less risky, and therefore increases the likelihood of successful technology transfer to industry [5]. In addition, in reliability experiments it is not the students' direct results that are of interest. Instead, it is the variation among their results that is relevant. It is therefore not automatically a problem if students achieve different results than professionals, as long as the experiment allows us to draw general conclusions about the method. In the lab (using students) and in the field (using experts) the participants to a reliability experiment should be as similar to each other as possible in background and experience.

Third, participants must *be sufficiently motivated* to apply the method to the best of their abilities. When tired or faced with a tedious and uninteresting task, the quality of the results will suffer. Experiments using students are sometimes conducted as part of a software engineering course. The experimentor then should consider whether compulsory participation offers sufficient motivation, or whether voluntary participation based on students' intrinsic motivation would be preferable. Furthermore, when the task at hand requires estimation (as will be the case for risk analysis), particular care should be given to avoid personal biases that can result in over- or underestimation. A frequently used way to control this bias is to employ teams instead of individuals. Discussion within the team can dampen individual biases.

b) Case to Which the Method is Applied. A method such as RASTER is not designed for a single case, but should perform well in a large variety of cases. If a case is ill-defined, then one cannot blame the method if it provides results with low reliability. Since testing of reliability requires a constructed laboratory experiment, the experimentor must carefully design the case to be used by the participants. The case should be representative of cases encountered in practice, but reduced to the essentials to make it fit to the limited time and setting available to laboratory experiments.

c) Environment During Application. Variation may also derive from *environmental conditions*, such as the meeting room, lighting, or time of day. First, the conditions should be as similar as possible between the different subjects to the experiment. Ideally, conditions should be identical. If the conditions differ, then any variation in results could be attributed to these conditions and does not necessarily indicate a lack of reliability. Secondly, the conditions should be as similar as possible between the experiment and real world applications of the method. For example, the experiment should, or should not be performed under pressure of time, depending on what is the case in practical applications. If the conditions differ, then it could be argued that variation in lab results would not occur in the field.

2) Validation of Effectiveness. When causes of contextual variation have been identified and mitigated, it is necessary to give a convincing argument that mitigation has been effective. The results of the method's application cannot be used in this argument, because the results may vary due to properties of the method rather than due to contextual factors. Instead, it will be necessary to make additional observations, using tools such as interviews, questionnaires, and observations. Therefore experiments to test reliability of a method will collect two kinds of data: measurements on the results of the method, and measurements on the usage of the method. We now discuss how we will analyse the results of the usage of the method.

3.2 Analysis of Measurements on the Results of the Method

The analysis of the reliability of a method can make use of several well-known statistical techniques for inter-rater reliability [6]. Inter-rater reliability is the amount of agreement between the scores of different subjects for the same set of items.

Well-known measures for inter-rater reliability are Cohen's kappa, Fleiss' kappa, Spearman's rho, Scott's pi and Krippendorff's alpha [15]. Cohen's kappa and Scott's pi are limited, in that they can only handle two raters. To test reliability, more than two outcomes are necessary in order to be able to draw conclusions. Fleiss' kappa can handle multiple raters but treats all data as nominal. Spearman's rho can take ordinality of data into account, but only works for two raters. Krippendorff's alpha works for any number of raters, and any type of scale. Furthermore, Krippendorff's alpha can accommodate partially incomplete data (e.g. when some raters have not rated some items). This makes Krippendorff's alpha a good choice for this domain. We will abbreviate 'Krippendorff's alpha' to 'alpha' in the remainder of this document.

Alpha is defined as $1 - D_o/D_e$, where D_o is the observed disagreement in the scores and D_e is the expected disagreement if raters assigned their scores randomly. If the raters have perfect agreement, the observed disagreement is 0 and alpha is 1. If the raters' scores are indistinguishable from random scores then $D_o = D_e$ and alpha is 0. If alpha < 0, then disagreement is larger than random disagreement. Alpha is therefore a measure for the amount of agreement that cannot be attributed to chance. Cohen's kappa and Scott's pi are basically defined in the same way as alpha, but differ in their computation of observed and expected (dis)agreement. For alpha and ordinal data, the disagreement between two scores s_1 and s_2 of an item depends on how often these scores occurred in the observed scores, as well as how often the levels between s_1 and s_2 have been used. The observed disagreement can thus be calculated by summation, and the expected disagreement can be computed based on the relative frequency of each ordinal level. More information is given in our internal report [21].

In order to compare the inter-rater reliability of two subsets of items, the calculations must be done carefully to ensure that the alphas are comparable. To be able to compare alphas on two subsets, we must ensure that the D_e value for both alphas is calculated over the complete set of scores, not over the their respective subsets [14]. In a subset the disagreement within the items

may be large when seen in isolation, while that disagreement may be much smaller when compared to the totality of scores. Since the absolute values for D_o and D_e depend on the number of items over which they are calculated, a scale factor must be applied during calculation. This computational complexity is not a peculiarity of alpha alone; by analogy, the other measures of inter-rater reliability are affected by a similar issue.

4 Research Method

4.1 Experiment Design

We conducted a replicated experiment in which small teams of volunteers performed part of the RASTER method on a fictitious telecom service. Detailed information on the practical aspects of the experiment are given in our internal report [21]. The following description follows the checklist given by Wieringa [25].

Treatment Design. Executing RASTER means that several activities need to be performed, ranging from collecting information to obtaining go-ahead from the executive sponsors of the risk assessment (see Fig. 1). Not all of these are relevant for reliability, because not all of them can contribute to variation in results. From previous research we know that different experts can create architecture diagrams in RASTER that are largely identical. We consider this part of the method reliable, and exclude it from our current experiment. In stage two, most of the expert assessments of frequencies and impacts of vulnerabilities are made, and so this stage is an important source of possible variation. Stages three and four add no other sources of variation. Including them in the experiment would greatly complicate the experiment without adding new knowledge about sources of variation, and so we decided to restrict our experiment to stage two.

Choice of Volunteers. To ensure sufficient knowledge of information technology (IT) security, we recruited student volunteers from the Kerckhoffs Masters programme on computer and information security offered jointly by the University of Twente, Eindhoven University of Technology, and Radboud University Nijmegen [13]. Since our groups are not random but self-selected, our experiment is a quasi-experiment [18]. This creates systematic effects on the outcome, that we will discuss in our analysis of the outcomes.

RASTER is applied by a team of experts. This allows for pooling of knowledge and stimulates discussion. We acquired 18 volunteers, enabling us to form 6 groups. Our sample of groups is not randomly selected and is anyway too small for statistical inference, but we will use similarity-based reasoning to draw tentative generalisations to analogous cases [1,4,19,25].

Target of Assessment. The telecom service for the experiment had to be small so that the task could be completed in single afternoon, but large enough to allow for realistic decisions and assessments. The choice of students imposed further restrictions; wireless telecommunication links had to be omitted (as students were unlikely to have sufficient knowledge on these), and a telecom service

was chosen to be relatively heavy on information technology. The telecom service for the experiment was an email service for a small fictitious design company heavily dependent on IT systems (Fig. 2).

Measurement Design. For measurement of the results of the method, we used the risk assessment scores by the groups. Groups were instructed to try to reach consensus on their scores. Each assessment was noted on a provided scoring form (one form per group). The possible scores form an ordinal scale. Detailed scoring instructions and descriptions of each of the values were included in the hand-out. In addition, groups could decide to abstain from assessment. Abstentions were allowed when the group could not reach consensus on their score, or when the group members agreed that information was insufficient to make a well-informed assessment.

For measurements on the usage of the method we used an exit questionnaire and our observations during the experiment. Each participant individually completed an exit questionnaire at the end of the experiment (Table 1).

4.2 Using our Approach to Testing Reliability

Subjects Applying the Method. Participants should understand the task, be able to perform the task, and be willing to do so. To mitigate lack of understanding, we provided a concise case which would be easy to explain; we prepared what we hoped were clear instructions and reference materials. We then tested the instructions (as well as the task itself) in a try-out. As a result of the try-out, we made small improvements to the instructions and to the case description. At the start of the experiment we invited questions from the participants and made sure to proceed only after all confirmed that they understood the task at hand. To mitigate lack of experience, we created a case that closely matched the expected experience of our students. As explained, we omitted wireless technologies, and emphasised IT systems in the case. To mitigate lack of motivation we recruited volunteers, offered the customary compensation, and raffled cinema tickets as a bonus.

Case to Which the Method is Applied. Two causes of variation drew our special concern. First, the *number of risk scenarios* could be too large. In the experiment, risks consist of the combination of an architectural component and a vulnerability, e.g. "power failure on the mail server". Many different scenarios can be devised for this risk to occur. For example, a power cable can be accidentally unplugged, the fans in the power supply unit may wear out and cause overheating, or the server can be switched off by a malicious engineer. A large number of risk scenarios will make the results overly dependent on the groups' ability to identify all relevant scenarios. Given the limited time available for the experiment, groups could not be expected to identify all possible ways in which a vulnerability could materialise. As a mitigation, we tried to offer clear and limited vulnerabilities in the case description.

Second, reliability cannot be achieved if there is *widespread disagreement on the 'true' risk* in society. Physical risk factors can in principle be assessed objectively, but some risk factors (such as fairness or voluntariness) are unavoidably

subjective. We therefore use quotation marks for 'true'; in some cases no single, most valid risk assessment may exist. Such controversial risks do not lend themselves to impartial assessment. In our choice of the experiment's case we tried to avoid controversial risks.

Environment during Application. We did not identify important causes of variation that needed mitigating. We provided each team with a quiet and comfortable meeting room, in a setting not unlike real world applications of RASTER.

Verification of effectiveness. For each source of external variation thus identified, the questionnaire checked whether participants had the required knowledge, ability and motivation to apply the corresponding countermeasure. For example, for 'lack of knowledge or experience' we used these three questions: "I am knowledgeable of the technology behind office email services", "My knowledge of the technology behind office email services could be applied in the exercise", and "It was important that my knowledge of email services was used by the group".

We also used the opportunity to include four questions to test some internal sources of variation. In particular, we wanted to test whether the scales defined for Frequency and Impact were suitable, and whether the procedure to avoid intuitive and potentially biased assessments was effective.

5 Results

Each of the six teams scored 138 Frequency assessments and 138 Impact assessments. Our scale for each is ⟨*extremely low, low, moderate, high, extremely high*⟩, but groups were also instructed that they could abstain from scoring. The experiment results can therefore be described as having 6 raters that had to rate 276 items, and partially incomplete, ordinal data. We computed alpha over these items, but also computed alpha over subsets of items. Subsets included the Frequency scores and Impacts scores separately, the scores on a single architectural component, and the scores on a single vulnerability.

Detailed results can be found in [21]; anonymised copies of results can be made available on request.

5.1 Scoring Results

Over the entire set of items, alpha is 0.338. This is considered a very weak reliability; Krippendorff recommends alpha at least 0.667 for provisional conclusions, and alpha at least 0.8 for firm conclusions, although he stresses that these figures are guidelines only [15]. Over the Frequency scores alpha is 0.232; over the Impact scores alpha is slightly higher, at 0.436.

This relatively low level of agreement is in line with earlier findings about reliability of risk assessments [7,8]. To understand why the level of agreement is relatively low we turn to the exit questionnaires.

Table 1. Exit questionnaire questions. Answers were scored on a five-point scale (strongly disagree to strongly agree, unless indicated otherwise).

1. The instructions at the start of the exercise were (very unclear – very clear).
2. I knew what I needed to do during the exercise.
3. In the experiment I could practically apply the instructions that were given at the start of the exercise.
4. The instructions that were given at the start of the exercise were (mostly useless – very useful).
5. My knowledge of the technology behind office email services can be described as (non-existent – excellent).
6. My knowledge of the technology behind office email services could be applied in the exercise.
7. It was important that my knowledge of email services was used by the group.
8. Before the exercise I was instructed to make rational, calculated estimates.
9. During the experiment I knew how to avoid fast, intuitive estimates.
10. The instructions and procedures for avoiding fast, intuitive estimates were (very cumbersome – very easy to use).
11. When estimating Frequencies and Impacts of vulnerabilities, it is necessary to consider many possible incidents.
12. I could think of practical examples for most of the vulnerabilities.
13. When discussing vulnerabilities, other members of my group often gave examples that I would never have thought of.
14. In my group we mostly had the same ideas on the values of estimates.
15. The estimates made by other groups (compared to ours) will be (very different – very similar).
16. For all estimates, there exists a single best value (whether we identified it or not).
17. I was able to concentrate on the exercise and work comfortably.
18. The time to complete the exercise was (way too short – more than sufficient).
19. Participating in this experiment was (very tiresome – very interesting).
20. The scales for values of Frequency and Impact estimates were (very unclear – very clear).
21. In my group we hesitated between two adjacent Frequency and Impact values (almost always – almost never).
22. The scales of values for Frequency and Impact were suitable to this exercise.
23. The final answer of my group often equalled my immediate personal estimate.

5.2 Exit Questionnaire

The exit questionnaire is shown in Table 1; answers are summarised in Table 2. The following discussion also makes use of our observations of the participants during the experiment.

Answers to the questionnaire (q1–q4) indicate that participants believe they had the required knowledge, skill, and motivation to employ the method. Our observations during the experiment confirm that, except for a few isolated cases, the instructions were effectively included in the groups' deliberations. We conclude that our mitigations for lack of understanding were successful.

The answers to q5–q7 were mostly positive, but our observations showed a marked difference in practical experience between groups. Some participants, contrary to our expectations, did not fully understand the function and significance of basic IT infrastructure such as DNS servers. To check whether lack of knowledge did induce variation in the scores, we compared the inter-group variation for components that are relatively well-known (such as desktop and

Table 2. Total scores for each of the questions in the exit questionnaire (see Table 1).

Question	*Strongly disagree*	*Disagree*	*Neither agree nor disagree*	*Agree*	*Strongly agree*
1				IIIIIII	IIIIIIIIII
2			I	IIIIIIIII	IIIIII
3			I	IIIIIIIIIII	IIII
4				IIIIIII	IIIIIIIIII
5		I	III	IIIIIIIIII	II
6			III	IIIIIIIIIII	II
7		III	IIIIIIII	IIIIII	
8		II	I	IIIIIIIIIII	I
9			IIII	IIIIIIIIIII	
10		II	IIII	IIIIIIIIII	
11		I	I	IIIIIIIIIII	II
12			II	IIIIIIIIIII	III
13		III	IIIIIII	IIIIIII	
14		I	II	IIIIIIIIIII	I
15		I	III	IIIIIIIIIII	I
16	IIII	IIIIIII	IIII	II	
17			III	IIIIII	IIIIIIII
18		IIIII	IIIIIII	III	II
19		II		IIIIIIIIIII	III
20		I	II	IIIIIIIIIII	
21	I	IIIIIIIII	IIIII	II	
22		II	IIIIIII	IIIIIII	
23		I	II	IIIIIIIIIII	I

laptops), to that for components that are less familiar (such as firewalls and routers). Alphas over end-user components (0.383 for frequencies assessments, 0.448 for impact assessments, 0.416 combined) were indeed higher than the general scores (0.232, 0.436, 0.338 respectively). We therefore conclude that lack of experience can explain some of the variation in results.

The answers to q8-q10 suggest that participants succeeded in avoiding personal biases. This was confirmed by our observations. We therefore conclude that our mitigations for lack of motivation were successful.

Two causes of variation arising from the case were identified and mitigated: a high number of risk scenarios and widespread disagreement on the 'true' risk. On the number of risk scenarios answers to the questionnaire (q11–q13) and observations indicated that mitigations were successful. In cases when the number of scenarios seemed unlimited (e.g. the risk of a general cable break in the Internet), groups did not hesitate to abstain from answering. For the second cause ("no 'true' risk") the questionnaire results were mixed: positive on agreement within the group and expected agreement with other groups (q14–q15), but negative on whether a single best assessment is possible (q16). The positive results could be a reflection of pleasant, cooperative teamwork, but the negative result to q16 makes it clear that participants believe there is no true answer. Our observations are that most groups made assumptions that significantly affected their assessments. The one group that scored high on q7 ("It was important

that my knowledge of email services was used by the group") also was the only group that scored positively on q16. This indicates that the participants probably recognised that their assumptions were somewhat arbitrary. The scoring forms had space for groups to mark important assumptions; none of these assumptions were extraordinary or unrealistic. We did observe that groups generally made many more assumptions than were noted on their forms, but these unrecorded assumptions were mostly natural or obvious. Based on the above, we conclude that variation in scores can be partly explained by the difference in assumptions made by groups.

Mitigation of causes of variation from environmental conditions appear to have been successful. Neither questionnaire (q17–q19) nor observations indicate that conditions affected the results unequally. One group finished within the time set for the task, others exceeded that time by a few minutes, although one group finished almost 45 minutes late. All groups completed their tasks.

To summarise, the measurements on the usage of the method indicate two unmitigated contextual causes for variation: participants' lack of experience and knowledge about IT systems, and different assumptions made by the groups. We now turn to sources of variation internal to the RASTER method itself. From these we discovered a third cause for variation.

The questionnaire (q20–q22) and our observations showed that groups often hesitated between two adjacent frequency or impact classes (recall that all assessments required the selection of a value from an ordinal scale). Participants also remarked that the range of the scales was large, and that the difference between adjacent steps was problematic. We observed that participants volunteered arguments pro and con, and referred to previous scores to ensure a consistent scoring. This was independent of the particular ordinal value; discussion was necessary for the extreme scores as well as for the moderate scores. It is likely that groups settled for different values in these discussions. A third, method-internal cause of variation in outcomes is therefore the difficulty in choosing between adjacent ordinal values.

5.3 Implications

We found three explanations for the variation in the assessments:

1. The lack of expert knowledge by the participants.
2. The difference in assumptions made by groups.
3. The need to make somewhat arbitrary choices between adjacent ordinal values.

In practical applications of RASTER the team of analysts would consist of industry professionals, and lack of knowledge (1) is therefore not expected. Also, in a field setting analysts have ways to deal with unavailable data other than making assumptions (2). For example, they can make additional observations, conduct inspections, actively look for further documentation, or interview actors to fill gaps in available information. The number and severity of assumptions would

therefore likely be much lower in field settings. In practice the team of analysts will be larger than the in the experiment (three students), allowing for more interaction and deliberation in order to reach consensus. Again, this suggests that in practice reliability of RASTER may be higher. These differences between lab and field suggest that RASTER will produce less variable results in practice than in the lab. Our experiment provides insight in why this can happen; only further field studies can demonstrate whether this is indeed the case.

However, explanation (3) will also be present in the field, and therefore is a point for improvement in the RASTER method.

6 Discussion, Conclusion, and Future Work

We have presented an approach to validating and measuring the reliability of a method, presented a research design that used this approach, and discussed the result of using this for RASTER. Our approach to measuring reliability of methods does not mention risk assessment at all, and should be of use also in measuring the reliability of other methods. The research design too should be of use for measuring other properties of methods.

Our analysis confirms that reliability of expert judgements of likelihood and impact is low. Our results add to this a quantification of that lack of reliability, using statistical tools, and a careful explanation of all possible sources of this lack of reliability, in the method as well as in its context of use. We conclude from our quantitative analysis that reliability in risk assessment with scarce data, that has to rely on expert judgement, may not be able to reach the standards common in content analysis. It may very well be the case that it is not achievable for any method that requires a very high amount of expert knowledge.

If this is true, then experts performing such assessments retain a large responsibility for their results. Their risk assessments not only yield risk evaluations, but also justifications for these evaluations, along with best guesses of likelihood and impact. They have to communicate these limitations of their evaluations to decision-makers who use their evaluations.

Based on our results, we have identified several ways in which the RASTER method could be improved. For example, RASTER currently defines medium impact as "Partial temporary unavailability of the service for some actors"; high impact as "Long-term, but eventually repairable unavailability of the service for all actors." Participants struggled with assessing risks that, for example, led to *partial* unavailability for *all* actors. We are currently working on improvements of the scales that do away with these ambiguities, which we will validate in a follow-up experiment.

Acknowledgments. The paper benefited from early comments by Sjoert Fleurke on inter-rater reliability, and Nazim Madhavji on research goals and research design.

References

1. Bartha, P.: By Parallel Reasoning. Oxford University Press (2010)
2. Feather, M.S., Cornford, S.L.: Quantitative risk-based requirements reasoning. Requirements Engineering **8**(4), 248–265 (2003)
3. Franqueira, V.N., Tun, T.T., Yu, Y., Wieringa, R., Nuseibeh, B.: Risk and argument: a risk-based argumentation method for practical security. In: 2011 19th IEEE International Requirements Engineering Conference (RE), pp. 239–248. IEEE (2011)
4. Ghaisas, S., Rose, P., Daneva, M., Sikkel, K.: Generalizing by similarity: Lessons learnt from industrial case studies. In: 1st International Workshop on Conducting Empirical Studies in Industry (CESI), pp. 37–42 (2013)
5. Gorschek, T., Garre, P., Larsson, S., Wohlin, C.: A model for technology transfer in practice. IEEE Software **23**(6), 88–95 (2006)
6. Hallgren, K.A.: Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods for Psychology **8**(1), 23 (2012)
7. Herrmann, A.: Information need of IT risk estimation - qualitative results from experiments. In: Proceedings of the REFSQ 2011 RePriCo Workshop, pp. 72–84 (2011)
8. Herrmann, A.: REFSQ 2011 live experiment about risk-based requirements prioritization: The influence of wording and metrics. In: Proceedings of REFSQ 2011, pp. 176–194 (2011)
9. Herrmann, A., Paech, B.: Practical challenges of requirements prioritization based on risk estimation. Empirical Software Engineering **14**(6), 644–684 (2009)
10. Höst, M., Regnell, B., Wohlin, C.: Using students as subjects–a comparative study of students and professionals in lead-time impact assessment. Empirical Software Engineering **5**(3), 201–214 (2000)
11. IEC: Analysis techniques for system reliability - Procedure for failure mode and effects analysis (FMEA). International Standard 60812:2006 (2006)
12. ISO: Risk management - principles and guidelines. International Standard 31000 (2009)
13. Kerckhoffs Institute: The Kerckhoffs masters programme. http://www.kerchoffs-institute.org/ Last accessed 2014–07-10
14. Krippendorff, K.: Calculation of alpha over partitions (Private communication)
15. Krippendorff, K.: Content analysis: an introduction to its methodology. 2nd edn. Sage Publications (2004)
16. Ojameruaye, B., Bahsoon, R.: Systematic elaboration of compliance requirements using compliance debt and portfolio theory. In: Salinesi, C., van de Weerd, I. (eds.) REFSQ 2014. LNCS, vol. 8396, pp. 152–167. Springer, Heidelberg (2014)
17. Runeson, P.: Using students as experiment subjects-an analysis on graduate and freshmen student data. In: Proceedings of the 7th International Conference on Empirical Assessment in Software Engineering.-Keele University, UK, pp. 95–102. Citeseer (2003)
18. Shadish, W., Cook, T., Campbell, D.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company (2002)
19. Sunstein, C.: On analogical reasoning. Harvard Law Review **106**, 741–790 (1993)
20. Svahnberg, M., Aurum, A., Wohlin, C.: Using students as subjects-an empirical evaluation. In: Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 288–290. ACM (2008)

21. Vriezekolk, E.: Testing reliability of Raster - report of experiment with Kerckhoffs students. Technical report, University of Twente (2014)
22. Vriezekolk, E., Wieringa, R., Etalle, S.: A new method to assess telecom service availability risks. In: Mendonca, D., Dugdale, J. (eds.) Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management ISCRAM2011 (2011)
23. Vriezekolk, E., Wieringa, R., Etalle, S.: Design and initial validation of the Raster method for telecom service availability risk assessment. In: Rothkrantz, L., Ristvej, J., Franco, Z. (eds.) Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management ISCRAM2012 (2012)
24. Vriezekolk, E.: Raster documentation website http://wwwhome.ewi.utwente.nl/~vriezekolk/Raster/
25. Wieringa, R.: Design Science Methodology for Information Systems and Software Engineering. Springer (2014)
26. Willner, P.: Methods for assessing the validity of animal models of human psychopathology. In: Boulton, A., Baker, G., Martin-Iverson, M. (eds.) Animal Models in Psychiatry, I, Neuromethods vol. 18, pp. 1–23. Humana Press (1991)
27. Yin, R.K.: Case study research: design and methods - fourth edition. Sage Publications (2009)