A Survey-based Study of the Mapping of System Properties to ISO/IEC 9126 Maintainability Characteristics

José Pedro Correia
Software Improvement Group
Amsterdam, The Netherlands
j.p.correia@sig.nl

Yiannis Kanellopoulos
Software Improvement Group
Amsterdam, The Netherlands
y.kanellopoulos@sig.nl

Joost Visser
Software Improvement Group
Amsterdam, The Netherlands
j.visser@sig.nl

Abstract

The ISO/IEC 9126 international standard for software product quality is a widely accepted reference for terminology regarding the multi-faceted concept of software product quality. Based on this standard, the Software Improvement Group has developed a pragmatic approach for measuring technical quality of software products. This quality model introduces another level below the hierarchy defined by ISO/IEC 9126, which consists of system properties such as volume, duplication, unit complexity and others. A mapping between system properties and ISO/IEC 9126 characteristics is defined in a binary fashion: a property either influences a characteristic or not. This mapping embodies consensus among three experts based, in an informal way, on their experience in software quality assessment.

We have conducted a survey-based experiment to study the mapping between system properties and quality characteristics. We used the Analytic Hierarchy Process as a formally structured method to elicit the relative importance of system properties and quality characteristics from a group of 22 software quality experts. We analyzed the results of the experiment with two objectives: (i) to validate the original binary mapping and (ii) to refine the mapping using the elicited relative weights.

1 Introduction

The ISO/IEC 9126 international standard [5] provides a terminological framework that distinguishes six main characteristics of internal and external software product quality. One of these main characteristics is maintainability, which is further subdivided into the subcharacteristics of analysability, changeability, stability, and testability.

Based on this standard, the Software Improvement Group (SIG) has developed a pragmatic model for measuring the maintainability of software products [4]. This quality model introduces another level below the framework defined by ISO/IEC 9126, consisting of system properties such as volume, duplication, and unit complexity. Based on their experience in software evaluation, the authors of the model defined a mapping between these system properties and the maintainability subcharacteristics in a binary fashion: a property either influences a characteristic or not.

In this paper, we explore the possibility of refining the binary mapping of the SIG quality model into a weighted mapping. For this purpose, the mapping problem is viewed as a multi-criteria decision problem. The Analytic Hierarchy Process (AHP) is a method for decision making that reduces such complex decisions to a series of pair-wise comparisons [10]. From the answers to such comparisons, the AHP method allows to calculate the weight for each factor in the overall decision problem, i.e. the weight of each property in the overall quality judgment.

In structured interviews, we elicited answers to pair-wise comparisons not from a single decision maker, but from a group of experts. These answers were first transformed into a set of relative weights for each expert and then combined into a group consensus on the weights to be used in the mapping. We compared the original binary mapping with the experimentally established weighted mapping in order to answer the following research questions:

Q1 Does the weighted mapping represent agreement among the experts?

Q2 How similar are the weighted and binary mappings?

Q3 Can the differences between the mappings be used to refine the quality model?

The answer to Q1 allows us to assess the usefulness of the elicited opinions for answering the subsequent questions. The answer to Q2 can be used to assess the validity of the
This paper is organized as follows. In Section 2 we provide background information on the SIG quality model and the AHP elicitation method. Section 3 describes the design of our experiment while Section 4 reports and analyzes the obtained results. Possible threats to validity are listed in Section 5 and a discussion of the significance of the experiment and lessons to be learned is provided in Section 6. A comparison to related work is made in Section 7. Finally, in Section 8 we summarize the answers to our research questions and we indicate possible avenues of future work.

2 Background

The SIG Quality Model The SIG has developed a layered model for measuring and rating the technical quality of a software system in terms of the quality characteristics of ISO/IEC 9126 [4]. The layered structure of the model is illustrated in Figure 1. Source code metrics are used to collect facts about a software system. The measured values are combined and aggregated to provide information on properties at the level of the entire system, which are then mapped into higher level appraisals that directly relate to the ISO/IEC 9126 standard.

The rating of system properties is first done separately for each different technology, and subsequently aggregated into a single technology-independent property rating by weighted average according to each technology’s relative volume in the system. Further details on this lower layer of the model can be found elsewhere [4].

Property ratings are mapped to ratings of subcharacteristics, following dependencies summarised in a matrix (see Figure 2, following a refinement [1] of the original mapping [4]). In this matrix, a cross (×) is placed whenever a property is deemed to have an important impact on a certain subcharacteristic. The subcharacteristic rating is obtained by averaging the ratings of the properties where a cross is present in the subcharacteristic’s line in the matrix. For example, changeability is considered to be affected by duplication, unit complexity, and inward coupling and its rating is computed by averaging the ratings obtained for those properties. Finally, all subcharacteristic ratings are averaged to provide the overall maintainability rating. The mapping employed by the model is binary, in the sense that no weights are applied but each system property (or subcharacteristic) simply does or does not influence a subcharacteristic (or main characteristic).

The Analytic Hierarchy Process AHP is a decision making technique that reduces a complex multi-criterion decision to a series of one-to-one comparisons [10]. It has a variety of applications in economics, business, social sciences, politics and other fields. For prioritizing software requirements, AHP has been favorably compared to other approaches [6].

AHP leverages the human ability to compare single properties of alternatives. From the result of a large number of one-to-one comparisons, a weight is synthesized for each criterion. For each comparison, a rank is elicited from the decision maker on a Likert scale (e.g. using values between 1 and 9) and these ranks are stored in a matrix. Then the comparison matrix is normalized and its eigenvalues are computed. These eigenvalues play the role of coefficients/weights when someone wants to evaluate the alternatives for the objectives under examination. Additionally, a consistency ratio is determined for the decision maker, to quantify the degree to which the elicited rankings respect transitivity. An inconsistency of 10% or less is deemed acceptable [10].

Compared to other approaches to multi-criteria decision making, AHP enjoys a number of advantages [6]. In particular, AHP is easy to use, yields reliable results, and is tolerant to inconsistencies or other errors in the elicited opinions. On the other hand, AHP is known to be problematic to be scaled up. More details about the strengths and weaknesses of AHP can be found elsewhere [10, 6].
### Maintainability subcharacteristics

<table>
<thead>
<tr>
<th>Analysability</th>
<th>How easy is to diagnose the system and/or to identify which parts of the system to modify?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Changeability</td>
<td>How easy is to carry out a modification (includes designing, coding and documenting changes)?</td>
</tr>
<tr>
<td>Stability</td>
<td>How well does the system avoid unexpected effects after a modification?</td>
</tr>
<tr>
<td>Testability</td>
<td>How easy is to validate the effects of the modification?</td>
</tr>
</tbody>
</table>

| System properties |
|-------------------|-------------------------------------------------------------------------------------|
| Volume            | Overall size of the system                                                          |
| Duplication       | Relative amount of code that has a clone somewhere else in the system                |
| Unit size         | Size of units                                                                         |
| Unit complexity    | Complexity of units                                                                   |
| Unit interfacing  | Intricacy of unit interfaces                                                          |
| Inward coupling   | Distribution of incoming calls                                                        |
| Outward coupling  | Distribution of outgoing calls                                                        |
| Exception handling| How thorough is the handling of exceptional situations (differentiation of exceptions, reporting) |
| Test quality      | How well is testing done (presence of unit testing, usage of test framework, amount of test cases) |

**Table 1. Definitions of terms provided to the subjects at the start of each interview.**

### 3 Design of the experiment

**Participants** To elicit the necessary weights, a group of 22 participants was asked to provide their opinion. Participants were recruited among SIG personnel who are familiar with software evaluation as performed during software risk assessments [3]. All participants are familiar with the ISO/IEC 9126 standard and all had previous exposure to the SIG quality model. Out of these participants: 13 have been working in IT for more than four years, another 6 have been working for between one to three years, and only 3 had less than a year’s working experience. Based on their role in the company, these participants were categorised as follows:

- **Software engineers (9)** who develop and operate tools for analyzing source code and interpret their results.
- **Software quality consultants (9)** who assess the quality and inherent risks of software systems and report on these to IT management.
- **Researchers (4)** of post-graduate level and higher that work in the field of software analysis.

**Data Collection** Since AHP was employed as a technique for eliciting the weights, the questionnaires consisted of five pairwise comparison matrices.

- One 4-by-4 matrix was used to elicit the relative weights of the subcharacteristics of maintainability.
- For each of the four subcharacteristics, a 9-by-9 matrix was used for eliciting the relative influence of the system properties.

Participants were asked to provide comparison scores for the upper triangle of each matrix (excluding the diagonal), as the values for the other triangle are their reciprocals. Furthermore, the consistency ratio of the pair-wise comparisons was calculated to assess the reliability of the comparisons performed by each participant in the experiment.

Using AHP to elicit weights involves comparing each unique pair of system properties (or maintainability subcharacteristics) to determine which of the two is of higher importance, and to what extent. The number of comparisons required for a matrix with \( n \) elements is \( n \times (n - 1)/2 \). In the case of the model used for this case study, the subjects were asked to perform comparisons for one matrix with 4 elements and four matrices with 9 elements, adding up to a total of 150 comparisons.

The participants used a Likert scale of 1 to 5 to evaluate the importance of the system properties and maintainability subcharacteristics. The interpretation of this scale was the following:

1. **Equal Importance**: The two properties (or subcharacteristics) contribute equally to a given subcharacteristic (or to overall maintainability).
2. **Moderate Importance**: Experience and judgment favor one property (or subcharacteristic) over another.
3. **Strong Importance**: Experience and judgment strongly favor one property (or subcharacteristic) over another.
4. **Very Strong Importance**: A subcharacteristic (or property) is favored very strongly over another.
5. **Extreme Importance**: A subcharacteristic (or property) is favored over another with the highest possible order of affirmation.

We used this 5-level instead of the 10-level scale that is typically used in AHP [10] in order not to overwhelm the participants with options and not to require overly detailed judgments from them.
Interviews Each interview started with an explanation by the interviewer to the participant of the meaning of all quality characteristics and system properties, using the definitions listed in Table 1. Also, the notion of unit, used in those definitions, was explained as the smallest piece of code that can be executed and tested individually (for Java or C# a unit is a method, for C it is a function or procedure).

Then, the participant was asked to provide his/her opinion about the importance of system properties and maintainability subcharacteristics. This involved at first the comparison of the unique pairs of the maintainability sub-characteristics and then of the system properties, using the scale explained previously. The participant was given the opportunity to revise his/her opinion based on the following information:

- The sum of the derived weights (which are percentages that add up to 100%).
- The consistency ratio of the comparison matrices (which, according to [10], should be lower than 10%).

Data Analysis The main focus of the experiment was to construct, based on the individual weights, a group consensus that would be compared with the existing model. For this reason we were primarily interested in measuring the central tendency of the data. In order to do that we had to select between the mode, the mean and the median. The first one is more appropriate for nominal data where there is only one value that is more common than all the others are [12]. Since the derived weights are ratios that span a continuous scale, we decided not to use this type of measure. We chose the median as the measurement of central tendency since, unlike the mean, it is robust [2]. This means that it is not as affected by the presence of outliers. For this case study, this is desirable since we want the determined consensus to reflect the opinion of a majority.

Thus, for each factor in our model, we determined the median of the weights given by all participants. Per row, these medians were subsequently normalized such that they would again add up to 100%. In this way, a well-formed mapping was obtained with relative weights rather than binary relations.

In order to evaluate the level of consensus achieved we needed to employ a measurement of dispersion. The median absolute deviation (MAD)\(^1\) was chosen since (as the median) it is a robust measurement. Also, unlike the interquartile range, it is sensitive to both variability around the center as to variability in the tails, while not giving undue weight to the latter [2].

Before applying the MAD, we rescaled the weights. In the way we applied AHP, the weights on the level of subcharacteristics-characteristics could in theory range from 6% to 63%, whereas for the other level they could range from 2% to 38%. In order to reason more clearly about the MAD and to make the values comparable we first performed a step of rescaling according to these theoretical limits. For example, a weight of 20% on the second level would be scaled to \((0.20 - 0.02)/0.36 = 50\%\).

Finally, we used visual inspection of the histograms of calculated weights per system property and quality subcharacteristic in order to detect situations where there are multiple concentrations of opinions.

Regarding the dispersion of weights, the following hypothetical scenarios can be distinguished:

**Perfect consensus** In this situation the weights obtained would be the same for every subject, thus yielding a MAD of 0%.

**Overall disagreement** In this situation the weights obtained would be dispersed over the whole range in a uniform distribution. The MAD for a uniform distribution is 25%.

**Polarized opinions** This situation would yield the highest possible MAD: 50%. This would happen when two equally-sized groups occupy the two extremes of the range with no values in between.

The dispersion information (histograms and MAD values) were interpreted in light of these hypothetical situations.

Before assessing central tendency and dispersion, data reduction was applied. For the purposes of our experiment we decided not to take into account opinions that were not sufficiently consistent, i.e. that did not respect transitive closure to a sufficient degree. This was assessed by the consistency ratio, a control mechanism provided by AHP as described in Section 2.

4 Results of the study

All interviews were conducted within a period of 3 weeks, with each interview lasting about 45 minutes. The result was a set of 5 comparison matrices per subject (110 matrices in total) with per matrix the derived weights as well as a consistency value. Applying data reduction as described previously, we excluded 10 of the 110 matrices on the grounds of having inconsistency values above 10%.

From the remaining matrices per subject we calculated the median values and normalized them to construct the weighted mapping represented in Figure 3. We also calculated MAD values (with respect to the non-normalized median values) and rescaled them, as depicted in Figure 4.

---

\(^1\)The MAD is defined as the median absolute deviation from the median, that is, taking per measurement the absolute differences from the median, the MAD is the median of those values. Note that the MAD was calculated relative to the non-normalized median.
5.1

Thus, there is agreement among the experts on 31 out of 40 relations in the weighted mapping. At the level of mapping sub-characteristics onto the maintainability characteristic, half of the relations are non-consensual. For this reason, the cases with low MAD values were found to correspond to two different situations:

- The values are grouped around the median, but with high dispersion;

- There are two separate concentrations of values, as in a bi-modal distribution.

For examples of these two situations see Figure 6. By visual inspection, we identified 3 cases of the first situation (overall disagreement) and 6 cases of the second situation (polarized opinions). For 2 of the latter cases the number of observations per mode was similar (thus similar to a symmetric bimodal distribution), whereas for the remaining 4 cases one of the concentrations had clearly a higher number of observations. We summarized these findings in Table 2. Note that on the level of sub-characteristics 2 out of 4 relations are non-consensual, while on the level of system properties, 7 out of 36 relations are non-consensual. Of these 7 non-consensual relations, 3 concern analysability.

Thus, there is agreement among the experts on 31 out of 40 relations in the weighted mapping.
son, we will not include this level in the answers to the remaining research questions. At the level of mapping system properties to sub-characteristics, two-thirds or more of the relations are consensual for each of the sub-characteristics, including for the least consensual, analysability. In the sequel, we will suppress the non-consensual relations at this level from our analysis.

Q2: How similar are the weighted and binary mappings? The binary mapping can be made comparable to the weighted mapping by regarding cells without cross as 0% weights and dividing weights equally among the crosses in each row.

After transforming crosses into weights this way, a direct comparison can be made between the two mappings, using the absolute differences as a measure of dissimilarity. The differences per mapping relation can be seen in Figure 7. For the mappings as a whole, the median absolute difference was 8.2%, with absolute differences ranging from 6.1% (influence of unit size on stability) to 23.0% (influence of inward coupling on changeability). To compare, the median absolute difference between the transformed binary mapping and a transformed uniform mapping (crosses everywhere) is 11.1%, with a similar range. Thus, the elicited mapping is closer to the original binary one than a uniform mapping.

This direct comparison does not take into account that weights obtained from the binary mapping only take extremal values, giving it a discrete character, while the elicited mapping has a more continuous distribution of weights. To compensate for this effect, we also applied a transformation to the elicited mapping, to make it discrete. We substituted each weight that was higher than the corresponding weight in the uniform mapping (11.1%) by a cross and the others with no cross. The differences between the two mappings obtained using this approach are represented in Figure 8. The 7 relations for which the experts did not reach consensus were kept out of consideration. The differences can be summarised as follows:

- 2 new crosses are introduced;
- 6 crosses present in the original mapping are not present in the new mapping;
- 21 relations remain unchanged;
- 7 relations are kept out of consideration due to lack of consensus.

Thus, the binary mapping and the discrete version of the weighted mapping are the same for 72% of the relations for which consensus exists.

Q3: Can the differences between the mappings be used to refine the quality model? For the mapping from sub-characteristics to maintainability, the elicited mapping suggests to increase the relative weight of testability, and to decrease the weight for stability. However, the consensus among experts may be regarded too low to warrant a change to the model.

For the mapping from system properties to subcharacteristics, consensus is stronger. Excluding the relations deemed non-consensual, the most important changes suggested by the elicited mapping are:
• for analyzability, decrease the weight of duplication and outward coupling;
• for changeability, decrease the weight of inward coupling;
• for stability, increase the weight of duplication and unit complexity and decrease the weight of unit interfacing;
• for testability, decrease the weight of unit size and outward coupling.

These suggestions for a direction of change of the mapping are supported by the consensus among experts. However, whole-sale substitution of the original mapping by the weighted mapping does not (yet) seem to be justified. Firstly, for 7 relations, no consensus was reached, with at least one non-consensual relation for each row. Secondly, even if consensus would have been established for all relations, rounding must be applied to the weights to reflect the variation among experts. For example, to justify rounding to entire percentage points, the MAD could be required to not exceed 1%, which amounts to 2.78% after rescaling. But in our current data none of the relations with consensus satisfies this requirement (see Figure 4). Thus, rounding to 5 or 10 percentage points should be considered instead.

5 Threats to validity

In this section we present some potential threats to validity, according to a classification scheme described in [12], and how we addressed them.

Conclusion validity These types of threats to validity concern the ability to draw the correct conclusion from the experiment. One possible risk is that the experiment is conducted differently for different subjects. To counter this we used the exact same form for every subject and carefully defined the procedure of the interview.

Internal validity Threats to internal validity concern factors that can affect the outcome of the experiment with respect to causality, without the researcher’s knowledge. One thing that comes into play here is the quality of the instruments used for the execution of the experiment, which in our case are the forms with the comparison matrices. These included some validation mechanisms (such as, for example, making sure each row’s calculated weights would add up to 100%) and were thoroughly tested before the interviews to minimize this potential threat. Furthermore, during the interviews the task of filling in the forms was left to the interviewer to make sure they were used correctly.

Another potential threat concerns the influence of subjects among each other, which in our case means that subjects could be influenced by other subjects’ opinions. This threat was minimized by conducting each interview separately and re-assuring each interviewee that we wanted to measure his/her personal opinion independently of the group’s opinions.

Finally, the fact that all experts had previous exposure to the binary model could shape their judgments. For this reason we didn’t allow the participants to have access to the model during the interviews, as we wanted to know their own opinions. It is interesting here to mention that some of the relations established in the original model were not matched with a strong consensus and that some alterations with strong consensus were identified.

Construct validity These types of threats concern the interpretation of the experiment as a correct reflection of what we are trying to study. Some of the potential threats are connected with the fact that, because subjects are consciously part of an experiment they could act differently. One such threat would be the concern of the participants that they are going to be evaluated and that their opinions will be used in order to judge their performance. In order to obtain the participants’ consent, we described the objectives of the experiment to them and assured them that the results concerning their personal opinion would remain confidential and would not be used for their evaluation.

Another risk that falls into this category is the experimenter’s expectancies biasing the results. The interviewers were aware of this and that no comments or judgements should be made from their side regarding the subjects’ opinions.

External validity Threats to external validity concern the generalization of the results obtained. In our case, a significant part of the company was selected for the experiment, and subjects were selected from three different groups (engineers, consultants and researchers). This makes our experiment generalizable to the context of our company, which was the scope of the case study. Further generalizations would imply selecting people from other contexts, such as, for example, academia or other companies.

6 Discussion

About the experiment Within the context of SIG, a software quality evaluation company, we would like to validate its existing model and seek if there is potential for refinement. Our perspective was that we wanted to know how a group of SIG employees who are familiar with the model, were perceiving the relations of its elements (i.e. the importance of system properties to maintainability subcharacteristics). That is, what do they think based on their experience is the importance of the system properties to subcharacteristics, and of the subcharacteristics to maintainability. For
this reason AHP was employed as a technique and a series of interviews using comparison matrices as questionnaires was conducted. These interviews with the experts were conducted within a period of 3 weeks, with each interview lasting about 45 minutes. As soon as all the data were collected, the matrices with an inconsistency ratio above 10% were excluded from the analysis (10 out of 110). Then the rest of the weighted mappings of the individual participants were combined into a weighted mapping for all experts by calculating the median weight across subjects for all mapping entries. Based on the derived results and as described in Section 4 there is a degree of consensus but not at the level that this would lead to apply any refinements on the existing model. We suspect the following reasons for this:

- The awareness of all the potential impacts in each of the relations may not was the same for all experts. In order to perform a comparison of, let’s say the impact of two properties for a given subcharacteristic, a subject must first be aware of all the potential impacts of each of the two properties on the subcharacteristic in order to aptly quantify the relative importance of the two properties. In case subjects are unaware of a potential impact of a property, it can lead to that property being neglected. In case different subjects are aware of different potential impacts of a property, this can lead to divided opinions regarding that property. To counter this, a more controlled experiment could be performed where descriptions of all the potential impacts would be given to the subjects.

- The sheer number of comparisons that the subjects were asked to perform may have led to the lack of concentration, even within the limited time-span of 45 minutes. Here, the order in which the comparisons were made became relevant since sometimes the latter comparisons were more neglected.

About the AHP It was easy for the participants to understand how AHP works and how they should complete the comparison matrices. However the scale from 1-9 proposed in [10] has some levels of intermediate values that the experts didn’t find useful as they preferred expressing their opinions using only the basic levels of equal, moderate, strong, very strong and extreme importance. For this reason we decided to employ the scale of 1-5 which uses only these levels. Another aspect of the AHP was the ability of checking the reasoning of the participants based on the consistency ratio. This helped us reducing the collected data and analyse the opinions that were following, at a certain degree, the transitive closure property. However, the participants found the whole process tiring as they had to repeat for five times the same task, that is to complete the comparison matrices. This task would be even more time consuming and tiring in the case that we had more than nine system properties that we would like to compare.

One of the disadvantages of this method is that it does not scale-up easily as if the number of system properties grows then the number of required pair-wise comparisons increases significantly [6]. The reason though for the selection of this technique in order to elicit the weights is that it is trustworthy since the amount of redundancy in the pairwise comparisons makes it fairly insensitive to judgemental errors [6]. In AHP also, the resulting weights are relative and based on a ratio scale, which allows for assessing the elicited weights.

7 Related work

In the context of ERP selection, Liang et al. [8] present a procedure that combines the ISO/IEC 9126 with a variant of AHP (fuzzy AHP). One of the steps involves the elicitation of weights through comparison questionnaires. The procedure was applied in two cases studies, involving 9 and 8 subjects, respectively. The method of combining opinions of several subjects is not explained. No analysis is made of variability among subjects or similarity to a given model.

Svanhberg et al. [11] combine ISO/IEC 9126 and AHP in an empirical study of the quality attributes of software architecture patterns. The subjects of the study were 8 academic researchers, most with industry experience. Agreement among experts is assessed with principal component analysis (unsuccessfully), use of sum of squares as distance measure, and visual inspection.

Kanellopoulos et al. [7] presented a method for performing mappings of low-level source code metrics to high-level ISO/IEC 9126 quality characteristics by various aggregation and weighting procedures. In this method they employed clustering to obtain quality profiles, and AHP to indicate the importance of source code measurements to system properties and to quality characteristics. In this work the derived weights were provided by a small group of researchers.

Finally, Narman et al. [9] developed a metamodel for enterprise architecture quality analysis. Extended influence diagrams were used in order to capture the definitions and the causal relationships of system quality attributes. The ISO/IEC-9126 standard was used as a frame of reference for this metamodel. The proposed metamodel was used in an industrial setting and the weights and the causal relationships have been defined by a decision analyst. The metamodel proved to carry a certain degree of uncertainty as either the sources of information lack credibility or it is very expensive to collect enough data to dispense with all uncertainty.
8 Conclusions

Contributions We have conducted an experiment with 22 experts in the field of software quality in order to explore validation and refinement of the SIG quality model. For each expert a series of weights was elicited though the AHP technique. We have shown that agreement between experts can be addressed by measuring their deviation from a consensus, where the consensus was established as the median. We validated the original model by quantifying its distance from the elicited consensus model. Finally, based on the differences between the two models, we identified points on which the original model can be refined.

Future work Future work extends into several directions. Firstly, the collected data allows further analysis. For example, we have not yet explored differences between groups of experts (e.g. engineers, consultants, researchers) or other factors such as years of experience. Secondly, we identified a number of threats to validity, that must be countered through a combination of further analysis and possibly further data collection. Finally, we intend to design and execute further experiments to support refinement of the model. The answers to Q3 indicated directions for refinement, but did not provide grounds for actual modifications. We intend to work with our group of experts to create a more sophisticated theory of the influences of system properties on quality characteristics. This theory can then be tested with a combination of several experimental techniques, including survey-based knowledge elicitation as presented here.

References